AFRL-RH-FS-TR-2017-0026

Flight Tasks and Metrics to Evaluate Laser Eye Protection in Flight Simulators

Thomas K. Kuyk
Peter A. Smith
Solangia Engler
Engility Corporation

Julie A. Lovell
Barry P. Goettl
711th Human Performance Wing
Airman Systems Directorate
Bioeffects Division
Optical Radiation Bioeffects Branch

July 2017
Interim Report for Sep 1, 2010 – July 31, 2017

Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Bioeffects Division
Optical Radiation Bioeffects Branch
JBSA Fort Sam Houston, Texas 78234

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

Qualified requestors may obtain copies of this report from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

"Flight Tasks and Metrics to Evaluate Laser Eye Protection in Flight Simulators"

(AFRL-RH-FS-TR- 2017 - 0026) has been reviewed and is approved for publication in accordance with assigned distribution statement.

SHORTER.PATRICK.D.1023156390

Digitally signed by SHORTER.PATRICK.D.1023156390
DN: c=US, o=U.S. Government, ou=DoD, ou=PKI, ou=USAF, cn=SHORTER.PATRICK.D.1023156390
Date: 2017.09.11 07:33:00 -05'00'

PATRICK SHORTER, Maj., USAF
Branch Chief
Optical Radiation Bioeffects Branch

MILLER.STEPHANIE.A.1230536283

Digitally signed by MILLER.STEPHANIE.A.1230536283
Date: 2017.12.03 13:15:11 -06'00'

STEPHANIE A. MILLER, DR-IV, DAF
Chief, Bioeffects Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE ***(DD-MM-YYYY)*** 07 -08-2017	**2. REPORT TYPE** Interim Technical Report	**3. DATES COVERED** ***(From - To)*** Sept. 1, 2010 – July 31, 2017

4. TITLE AND SUBTITLE

Flight Tasks and Metrics to Evaluate Laser Eye Protection in Flight Simulators

5a. CONTRACT NUMBER
FA8650-14-D-6519

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kuyk, Thomas K.; Smith, Peter A.; Engler, Solangia; Lovell, Julie A.; Goettl, Barry P.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H0PG

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Bioeffects Division
Optical Radiation Bioeffects Branch

Engility Corporation
4141 Petroleum Rd
Fort Sam Houston
TX 78234-2644

8. PERFORMING ORGANIZATION REPORT

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Bioeffects Division
Optical Radiation Bioeffects Branch
JBSA Fort Sam Houston, TX 78234-2644

10. SPONSOR/MONITOR'S ACRONYM(S)
711 HPW/RHDO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-RH-FS-TR-2017-0026

12. DISTRIBUTION / AVAILABILITY STATEMENT Distribution, A: Approved for public release; distribution unlimited. PA Case No: TSRL-PA-2017-0250

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Testing laser eye protection (LEP) devices intended for aircrew use in flight simulators allows for experimental conditions that directly address questions regarding their effects on flight performance that cannot be achieved with laboratory vision tests and provides a comparable platform that replaces the prohibitively expensive dedicated flight tests. To determine what flight tasks and performance metrics might be the best for LEP assessment, a literature search was conducted. From this search, a set of papers representative of different uses of flight simulators to address flight performance was identified along with a small set of reports that investigated the link between specific aspects of visual function and flying performance. The papers were reviewed and annotated bibliographies prepared. The reviews were organized by topic areas and are described in this report along with suggestions about how to approach LEP evaluations in flight simulators in general, with some specific suggestions aimed for use of the flight simulator within the RHDO vision laboratory. The challenges of simulator testing are discussed. In addition, suggestions are made regarding experimental approaches associated with the goal of creating a comprehensive model of visual function as it relates to flying performance.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF: Unclassified			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	Unclassified	56	Dr. Barry Goettl
U	U	U			**19b. TELEPHONE NUMBER** ***(include area code)*** NA

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18

This Page Intentionally Left Blank

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AB	Annotated Bibliography
ADI	Attitude Direction Indicator
AGL	Above Ground Level
AFRL	Air Force Research Laboratory
AOI	Area of Interest
ASL	Applied Sciences Laboratory
CSF	Contrast Sensitivity Function
DA	Decision Altitude
EOG	Electro-oculogram
FMA	Flight Management Annunciator
HMD	Head Mounted Display
HUD	Heads Up Display
IMC	Instrument Meteorological Conditions (used interchangeably with IFR)
IFR	Instrument Flight Rules
LED	Light Emitting Diode
LEP	Laser Eye Protection
MAPP	Model Assessing Pilot Performance
OD	Optical Density
OTW	Out-The-Window
NIR	Near Infrared
NVG	Night Vision Goggle
PAPI	Precision Approach Path Indicator
PFD	Primary Flight Displays
PLT	Photopic Luminous Transmission
RGB	Red Green Blue
RHDO	711th Human Performance Wing, Airman Systems Directorate, Bioeffects Division, Optical Radiation Bioeffects Branch
RMSE	Root Mean Squared Error
RXAP	Materials and Manufacturing Directorate, Functional Materials Division, Photonic Materials Branch
SA	Situation Awareness
SAGAT	Situation Awareness Global Assessment Test

VASI	Visual Approach Slope Indicator
VMC	Visual Meteorological Conditions (used interchangeably with VFR)
VFR	Visual Flight Rules

This Page Intentionally Left Blank

1 OVERVIEW

Wearing laser eye protection (LEP) devices can change the appearance of colored stimuli, reduce overall light levels, and alter contrast levels between objects in a scene. As a result, LEP has the potential to impair flight performance by adversely affecting the visual abilities of aircrew. The effect could cause aircrew to find LEP unacceptable and possibly dangerous to wear during flight. Therefore, it is critical that LEP devices be evaluated thoroughly, preferably before they are handed to pilots for evaluation in ground and flight-testing. The 711th Human Performance Wing (711HPW), Airman Systems Directorate, Bioeffects Division, Optical Radiation Bioeffects Branch (RHDO) provides test support for evaluation and quantification of effects involving visual performance and cockpit compatibility of prototype aircrew LEP. The current evaluation paradigm is a four-phase process that begins with physical and optical quality testing (Brockmeier, Kuyk, & Novar, 2014), and investigations of the LEP effects on basic visual functions (Maier et al., 2007). If the results of these tests are acceptable, the LEP can progress to ground and flight testing. Aircrew evaluations of LEP compatibility with general flight operations, cockpit displays, instruments, indicator lights and maps, and the out-of-cockpit scene as well as a judgement of safe-to-fly (Novar et al., 2015).

However, the current LEP testing process does not provide quantitative data about performance on tasks such as target detection and tracking, or flying and landing under degraded conditions, to name just a few. This information, as well as much of that gathered from ground and flight testing, could potentially be obtained using a flight simulation environment. If simulator testing would occur before ground and flight-testing, it could be used to determine whether a prototype LEP is ready for these final two phases of testing as well as augment findings on LEP visual effects obtained in the other phases of evaluation. This report provides a review of literature on the use of flight simulators to assess pilot performance with respect to specific flight tasks and in some cases the relationship with distinct aspects of visual function. The discussion reports published results acquired by online literature searches that identify measures of pilot performance as well as methods and techniques used to assess them that are important to safe and effective aircraft piloting. The reason for identifying these tasks, performance metrics, and study designs was to determine those that might be affected by changes in the visual environment caused by wearing LEP and then use them to assess the impact of wearing LEP in a flight simulator environment.

2 INTRODUCTION

Vision is undoubtedly the most import of the senses for most tasks involved in flying an aircraft. However, studies that have directly examined the link between visual function and performance involving specific flight related tasks are rare, and a comprehensive model of visual function as it relates to flight performance does not exist (Kumagai, Williams, & Kline, 2005). LEP has the potential to alter distinct characteristics of the visual environment, giving rise to concerns over the impact on flight tasks and performance. Nonetheless, despite the lack of a model linking visual function to flight performance, flight and flight-related tasks can still be used to evaluate the effects of LEP on performance.

Laser eye protection works by reducing the amount of laser light reaching the eye to levels that would not cause damage or visual distraction. It does this either by absorbing, reflecting, or diffusing light at specific laser wavelengths. For LEP that block laser radiation at wavelengths visible to the human eye, this results in a change in the spectral content of light available for vision. The consequences can be alterations in the appearance of colored objects, changes in contrast and an overall tinting and darkening of the scene. These changes can have an adverse effect on the wearer's ability to detect and identify visual stimuli and in a flight situation can result in difficulty acquiring visual information from both inside and outside the cockpit. It is critically important to assess the visual impact of LEP during the development stage to insure these systems are compatible with the operational tasks to be performed while wearing them.

The current test paradigm used by the United States Air Force (USAF) to assess prototype LEP is a four -phase process that involves laboratory physical and optical quality testing, laboratory visual function testing, followed by ground testing and flight-testing with aircrew in representative aircraft (Kuyk et al., 2016; Putnam, Novar, et al., 2017). The ultimate purpose of the testing is to determine if a specific prototype LEP is safe to fly in a set of representative aircraft under daytime or nighttime conditions and can progress to the acquisition phase. During the acquisition phase, the test results for prototypes are used to guide development requirements and modifications to the filter design. Evolution in evaluation approaches also are created for best LEP testing strategies. However, throughout the course of prototype testing, the results can be applied to guide LEP design involving filter modifications to improve specific aspects of performance (Putnam, Goettl, Novar, Kuyk, & Smith, 2017).

The physical and optical quality testing determines if the LEP meets protection (optical density) and optical (haze, distortion and optical power) requirements. The visual function tests include contrast acuity and spatial contrast sensitivity with and without a glare source present, color discrimination, and color identification. Significant failures to meet protection and optical requirements can result in cancellation of the visual function testing and force modification of the filter protective design characteristics to meet requirements. Similarly, significant adverse visual effects can be found during laboratory visual function testing that may cause sufficient concern to stop the LEP advancing further in the evaluation process. However, the visual function tests provide only basic information about effects on vision and observed failures require content of extreme concern to preclude ground testing.

In the ground testing phase, pilots and aircrew sit in a powered-up, but stationary, aircraft. Using a rating scale, they evaluate the visibility and usability of the displays, gauges, and indicator lights inside the cockpit, the general out-of-cockpit scene and provide a judgment on the safety of the

LEP for use in flight. The results of the ground test are used by a safety review board in their approval process for LEP flight tests and for determining if any restrictions need to be placed on their use in flight environments. For example, a recent ground test found that the LEP under test made interpretation of symbology on the tactical display difficult and resulted in the safety review board imposing a restriction on the use of the LEP during the tactical phase of the flight (Novar et. al., 2015). During the flight test phase, pilots and aircrew evaluate the visual compatibility of the LEP with specific aspects of flight, form, fit and function, the visibility and usability of specific displays and instruments and the out-of-cockpit scene. Lastly, they provide a safe-to-fly rating.

The laboratory visual function testing currently does not include tests of complex and operationally relevant task performance. Ground and flight tests are restricted to a qualitative observational assessment of the visual compatibility of the LEP with displays, instruments and indicator lights with LEP, and provide no quantitative performance metrics. In addition, flight tests are performed on a tag-along basis, often without prior knowledge about the nature of the sorties during which the LEP are to be evaluated; this limits the information that can be obtained because the assessment cannot be tailored to a specific scenario. Moreover, it does not allow for quantitative assessment of performance on specific tasks such as: detection of air and ground targets, responses to warning lights, changes in display symbology, airspace aircraft tracking, landing, and flight in difficult and demanding conditions.

The most appropriate way to test LEP for visual compatibility would be to determine the effects on performance using tasks that are important to safe and effective flight and commonly performed by pilots, and to test these in an actual flight environment, with appropriate instrumentation to support quantitative metrics. This testing paradigm would require dedicated flights, which are expensive, time consuming, and difficult to coordinate. However, many of these issues could be addressed by testing LEP in flight simulators prior to testing in real aircraft. To perform an accurate form, fit, and functional assessment of LEP in flight simulators, the relevant tasks and performance metrics need to be identified. To provide initial support of this baseline informational need, a literature search was conducted to identify representative studies that assessed various aspects of flight performance and related tasks. The relevant literature was summarized in the form of an annotated bibliography, and the results organized and described in this document. The end goal is to use this information to guide development of methodologies to assess the impact of wearing LEP on flight performance in a flight simulation environment. The primary interest is creating metrics that can be implemented in the RHDO simulator as well as for applying protocols of broader interest for advancing simulator use in general.

Another area of interest where testing in simulators would be useful is to help determine what aspects of visual function are most important for performance of different flight tasks. This information could be used to develop a model of visual function as it relates to flying performance as well as new laboratory tests to better determine if a specific LEP device is worth testing further in simulator as well as in a real aircraft.

3 METHODS

The literature search was conducted on-line using different sets of key words including: aircraft, pilot, performance, simulator, flight simulator, visual function, scan patterns, gaze, cockpit instruments, cockpit displays, flight performance, attention, visual attention, pilot assessment, vision tests, visual perception, etc. The papers, presentations, and reports found covered a range of topics, not all of which were relevant to identifying flight tasks and metrics. A process was adopted whereby each captured article was reviewed to determine if it contained significant and relevant information. If a publication was deemed sufficiently relevant, an annotated bibliography for that paper was written. An example of an annotated bibliography is shown in Table 1. The format was adapted from Adams et al. (2013) and was selected for its simple and organized structure. Its tabular layout makes it simple for the reader to pinpoint desired information. Approximately fifty informational documents (e.g., journal publications, technical reports, and research presentations) are outlined in the annotated bibliography. Each annotation includes the following subsections: A general statement of the finding, Full Reference, Purpose of the Study, Method or Measure, Scientific Quality, Core Findings, Type of Sample/Pilot, Relevance, and Key Words.

For this literature review the papers were grouped into three major areas based on primary outcome measure/general methodology: eye movement studies, simulator performance studies, and visual sensory and perceptual function studies. Within the first two major areas there are sub-categories and there is some overlap in key areas. For example, studies that used eye movements as a dependent measure and studies where simulator flight performance factors were the dependent variables have both been used to evaluate displays.

The focus of this review is to summarize the methods used to evaluate performance on flying tasks in flight simulators. The summary is followed by an overview of LEP research as it relates to effects on visual function and perception. The last section presents recommendations on possible test scenarios to evaluate LEP using the RHDO flight simulator laboratory as well as how the simulator environment could be used to determine which aspects of visual function are important contributors to flight performance.

Table 1. Sample Annotated bibliography: Pilot performance quantified during IMC conditions

Full Reference: Crognale, M. A., & Krebs, W. H. (2008). Proceedings of the Human Factors and Ergonomics Society Annual Meeting: *Helicopter Pilot performance: inadvertent flight into instrument meteorological conditions*, Las Angeles, CA: Sage Publications.	
Purpose of the Study: Authors aimed to develop methods to quantify helicopter pilot effort and performance and applied these to simulated flight instrument meteorological conditions (IMC).	
Method/Measure: Twenty commercial instrument-rated pilots were tested in a helicopter simulator (Flyit) running a Microsoft flight simulation package for the Bell 206. The simulation scenery environments mimicked common commercial helicopter operations, including departure from an offshore site with a short flight to the "mainland" and a scenario flying along moderately mountainous and forested terrain. A flight instructor acted as an air traffic control for the pilots. After a few minutes of familiarizing with the flight characteristics of the simulator, each pilot flew 5 assigned "missions" at given altitudes and airspeeds (order was randomly chosen to avoid practice effects). During the presented scenario, pilots flew for 5 minutes, and then gradually viewed a reduction in visibility over a 3 minutes period until it hit 0% visibility. The pilot then had to make any decision he deemed best, for up to 15 minutes after, while being forced into IMC rather than allowed to descend below clouds and land. Data were collected from the simulator's computer program regarding the flight instruments, aircraft performance, scenario information, and control inputs. Data was recorded using a program that ran simultaneously with the simulator, downloading data and flight parameters in real time. Data could be played back using a different program (not named). Each of eight chosen measures (pitch power, pitch error rate, fore/aft cyclic power, bank power, bank error rate, lateral cyclic movement power, pedal movement power, and vertical airspeed error rate) were analyzed independently, using: participant, time in type, order of test, visibility, altitude, and speed as parameters in the model. The first analysis treats pilot inputs and aircraft performance parameters as time series data and utilizes the power in the "Fourier domain" as a dependent measure. The second analysis (error analysis) looked at pilot performance by comparing the "error" rate that occurred during the VFR and IFR portions of the flight. Data from the power analysis (Fourier domain) were tested using a general linear model from SAS. The data from the error analysis contained many zeros because there were times that the pilots made no errors. The model we chose for the error analysis was the generalized linear mixed model as provided by SAS. The reported probabilities (*p*) are from the chi-square of the differences of the least-squares mean.	
Scientific Quality: Peer-reviewed. Sizable N (20 medium to high-time commercial pilots) due to large interest in study. Detailed statistical analysis. High-quality.	
Core Findings: Degradation of pilot performance occurs during encounters with inadvertent IMC condition; although; many of the pilots improved dramatically with increased testing suggesting that short training periods would greatly improve performance during inadvertent IMC encounters. Both the power analysis and error analysis provide valuable information regarding pilot performance. The power analysis was particularly helpful as an objective and continuous measure of pilot control inputs. The analysis revealed that cyclic inputs, but not rudder inputs, are a sensitive indicator of pilot workload or effort.	
Type of sample/pilot: commercial pilots	*Relevance:* The methods developed for quantifying pilot effort and performance can be applied to wide range studies of pilot performance. The analysis of pilot simulator control inputs is used as a metric for objectively quantifying pilot effort or "workload".

Key Words: helicopter, performance, inadvertent flight, meteorological, pilot, simulator, IMC, Flyit, Microsoft flight, terrain, instructor, ATC, altitude, airspeed, reduction in visibility, visibility, decision, aircraft performance, scenario information, control inputs, pitch power, pitch error, cyclic power, bank, bank, lateral cyclic movement, pedal movement, and vertical airspeed, error rate, speed, error analysis, pilot performance, general linear model, training, rudder

4 RESULTS

Almost every study reviewed involved testing in flight simulators. There are a few exceptions, but those selected for inclusion used isolated, but relevant, components of cockpit displays. The simulators used for testing ranged from simple desktop systems with a single display and limited controls to sophisticated full-motion systems representative of a specific airframe (Figure 1). Experimental subjects ranged from complete novices/university students to highly experienced commercial and military pilots. Subject age was also varied in some studies, to allow for the inclusion of analyses of age differences in the investigation.

(source: http://www.whiteman.af.mil/News/Features/Display/Article/326051/ and https://commons.wikimedia.org/wiki/File:USMC-02023.jpg Aug 2017

(Source: https://commons.wikimedia.org/wiki/File:AC97-0295-13_a.jpeg, Aug 2017)

Figure 1. Three flight simulators varying in level of sophistication and fidelity.

4.1 Flight Tasks

The studies reviewed initially can be linked to specific tasks that have been identified as performed by all pilots and a second list of tasks specific to air combat operations. A listing of these tasks is provided below for reference. The initial basis of the listing was a task analysis conducted by Kumagai et al. (2005) with military pilots of fixed and rotary wing aircraft who served as the focus group. They identified a set of tasks that are performed by all pilots regardless of aircraft type and that rely on the visual system. A small number of tasks on the Kumagai list were not performed in any of the studies reviewed and these are marked with asterisks. The list provided below was also augmented by additional tasks found in the studies reviewed, that were not on the initial list. The tasks that came from Kumagai et al. (2005) are listed first, with tasks taken from other studies listed at the end in italics. The common tasks identified were:

1. Reading VFR maps, contour maps, and approach plates
2. Interaction with and comprehension of instrumentation, displays, the weather radar display and the flight management system
3. Target/object detection and identification (e.g., landmarks, tower/ground signal lights, runway hazards, aircraft, ships, etc.).
4. Aircraft landing and take-off
5. Perceiving and responding to cockpit warning lights
6. Detecting other aircraft and/or birds in the peripheral field of view
7. Determining attitude (pitch, roll and yaw) of the aircraft
8. *Flying directly into the sun or flying with the sun directly behind the aircraft
9. Distinguishing differences between color coded items, both inside and outside the cockpit
10. Determining clearance distance between the aircraft and surrounding objects
11. Movement detection from within the aircraft cockpit (e.g., flashing warning lights)
12. Continuous transitioning between near and far viewing (cockpit instrumentation to exterior environment)
13. Low-level flying over various terrain
14. *Fast visual accommodation from bright light to low light and vice versa
15. Reading emergency checklists while flying and responding appropriately
16. *Executing heading and altitude changes*
17. *Avoiding air traffic*
18. *Taxing*
19. *Maintain heading and altitude*
20. *Determining heading direction and orientation of other aircraft*

In addition to the common tasks, several military specific flight tasks were also identified by Kumagai et al. (2005). Some of these fall under the general tasks listed above but for a specific object/target, such as smoke identification associated with search and rescue operations, and are not listed. Other military specific flight tasks not likely to be flown by civilian or commercial aviators and are listed below. Again, tasks on the list that were not seen in any of the studies reviewed are marked with asterisks, and tasks that were added to the list from other studies are listed at the end in italics. Note that most tasks, including those falling into the common task

category are performed day or night, with or without night vision goggles (NVG), and in VFR or IFR conditions.

1. Nap of the earth flight over smooth and rough terrain
2. Formation flight
3. *Tactical approach (steep glide path, high speed, evasive maneuvers)
4. *Glide path approach night unaided single light (rotary wing)
5. Approach and landing night unaided
6. *Hover/sling rescue (rotary wing)
7. Air refueling
8. Target detection through Heads up Display (HUD) or Helmet Mounted Display (HMD)
9. Aircraft or ship identification
10. *Target tracking, aiming, weapons release*

4.2 Eye Movement Studies

4.2.1 Distribution of attention/situation awareness.

Studies in this group used metrics derived from eye movement data as the primary dependent measures to assess distribution of attention and its relationship to situation awareness (SA). Where gaze is directed is generally considered to be an indicator of where attention is directed, and the information obtained during a fixation a precursor to cognitive processing. The premise is that individuals spend more time looking at important or interesting items. The duration of fixations is an indicator of the degree of difficulty extracting information and the number of fixations made to a specific item indicates its importance (Ahlstrom & Friedman-Berg, 2006; Yu, Wang, Li, Braithwaite, & Greaves, 2016). In a three phase theoretical model of SA proposed by (Endsley, 1995), eye movements have been hypothesized to be indicators of the first phase involving perception of cues (Biella, 2009; Yu, Wang, Li, & Braithwaite, 2014; Yu et al., 2016).

Moacdieh and Sarter (2012), in a study of the effects of clutter level on detection, investigated a number of eye tracking metrics that the prior literature had indicated are sensitive to the effects of clutter. The details of their experiment will be reviewed later with other studies that looked at clutter in displays. However, he list of metrics in their Table 1 provides a useful tool relevant to all the studies reviewed here, since these studies frequently apply one or more of these metrics and often for purposes other than evaluating the effects of clutter. It is important to note that most metrics listed are referenced to an area of interest (AOI). An area of interest is a pre-defined region in the cockpit space such as an individual instrument like the primary flight display (PFD), a smaller part of a display, a control such as the thrust lever, or the outside view, usually termed out-the-window (OTW). All the studies reviewed analyzed eye movement data relative to AOI's. Table 1lists the eye movement metrics, a description of what they are, what changes in them represent, the studies that used them (to evaluate clutter), and, finally, whether they were affected by clutter significantly in their experiment. The eye movement metrics listed include: total fixation number, cumulative fixations on a target AOI, number or percentage of fixations in an AOI, time

between first fixation in an AOI and finding the target, backtrack rate, gaze number in an AOI, mean gaze duration in an AOI, time to first fixation on an AOI, mean fixation duration in an AOI, scan path-length, mean saccade length, convex hull area, spatial density, and ratio of transitions.

The general format of the studies that involved eye tracking was to instrument a subject or the cockpit with an eye tracking system and have them perform different flight related tasks in the simulator (Figure 2). All the studies used the concept of AOI's and it can generally be said that AOI's included the primary flight instruments (usually the PFD) and OTW view. Some studies also included other displays and instruments, specific controls or a pre-defined part of a display as AOI's. The number of AOI's varied between studies, and depended on the type of aircraft simulated, the sophistication of the simulator, limitations of the eye tracker, and the goals of the study.

Figure 2. Subject equipped with iViewX (Sensomotoric Instruments) eye tracking system sitting in an Airbus style simulator (Biella, 2009)

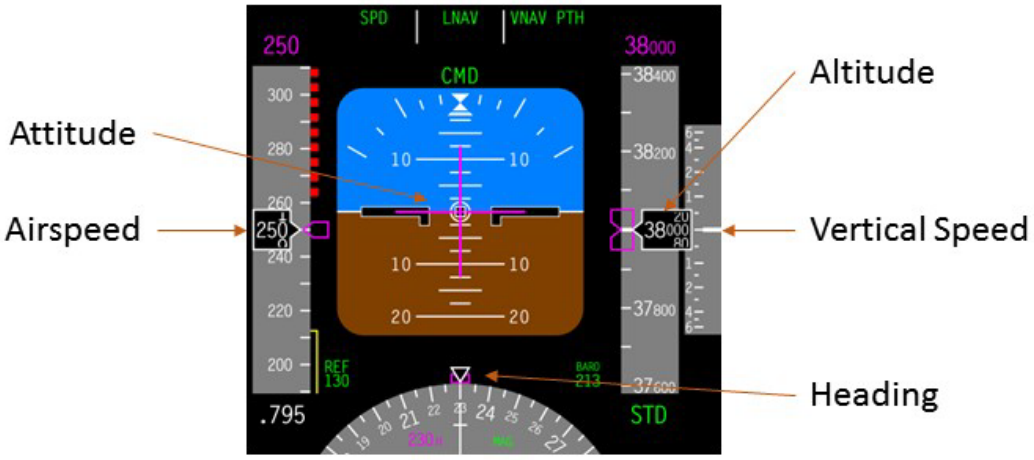
The most commonly used eye movement metrics included cumulative dwell times usually expressed as a proportion of time spent in each AOI, number of fixations on an AOI and mean fixation duration. As a caution, some of the terminology is ambiguous and used in diverse ways by different investigators. Although some metrics are obvious, like number of fixations, fixation duration, and number, velocity and length of saccades, others such as gaze point and gaze duration are not always defined well enough to differentiate them from fixation and fixation duration. All of the studies that looked at distinct phases of flight found they had different scan patterns, with scan pattern defined as how much time was spent looking at each of a set of AOI's. Similarly, studies that investigated experienced and inexperienced pilots usually found some differences in scan patterns for the two groups. Given these generalities, the review of the eye movement studies will focus on their purpose, the flight tasks and how performance was measured, the eye movement metrics used and how defined, general findings and limitations.

Since the PFD is a common AOI, or divided into multiple AOI's, some explanation of this instrument is useful. Prior to the 1980's, the main flight instruments (altimeter, air speed indicator,

vertical speed indicator, and attitude, heading, and turn indicators were commonly arranged in a "six-pack" formation (Figure 3). Modern cockpits combine this information in a single display, the PFD which comprises five primary information areas (Figure 4), and is usually presented on a cathode ray tube or liquid-crystal display device. PFD's, however, are not standardized, and vary considerably in complexity (Figure 5 and Figure 6).

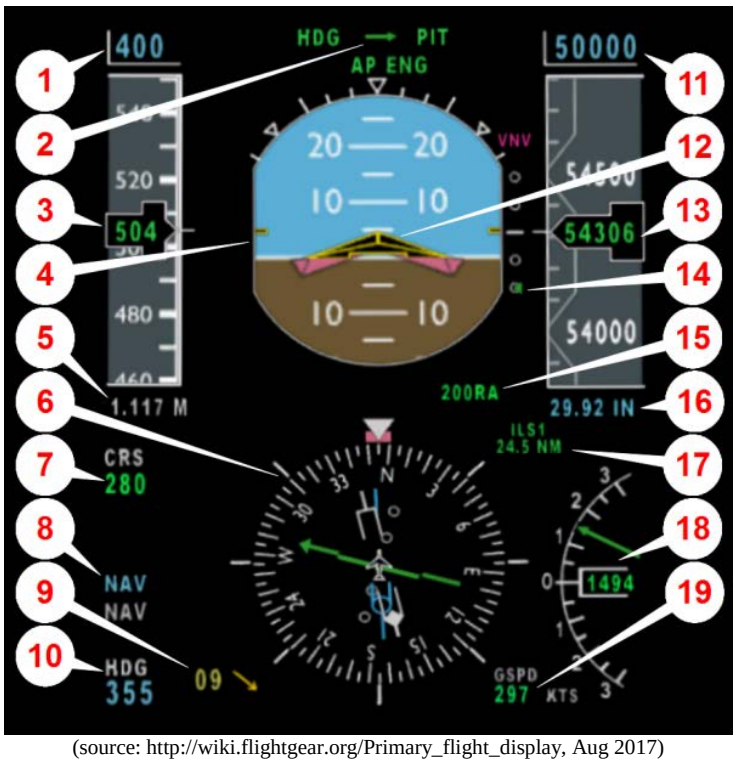
(source: https://commons.wikimedia.org/wiki/File:Six_Pack_flight_instruments.jpg, Aug 2017)

Figure 3. Example of the instrument "six-pack" of older aircraft



(source: https://commons.wikimedia.org/wiki/File:Primary_Flight_Display.svg, Aug 2017)

Figure 4. Simple PFD (left) with the five primary information areas



(source: http://wiki.flightgear.org/Primary_flight_display, Aug 2017)

1. Autopilot speed, sync with actual or set with autopilot	11. Autopilot altitude, change with MFD knob, sync with actual or set with autopilot
2. Autopilot mode	12. Single cue or cross pointer (cross pointer is not allowed in Europe)
3. Indicated airspeed	13. Indicated altitude. From 500 feet AGL and lower the "ladder" will also show ground
4. Attitude indicator	14. Glideslope pointer (dot above center = flying below glideslope)
5. Indicated airspeed in Mach (yes, this is a combined image but the values were real)	15. Decision height (can only be set on ground)
6. Double bearing HSI (blue / white) with course deviation bar (green) and heading bug (red)	16. Altimeter correction setting, QNH
7. Set radial of selected NAV, change with standby HSI or Autopilot	17. Distance to NAV1 or NAV2, indication of NAV type (VOR, ILS, FMS)
8. Bearing type (NAV, ADF, FMS) in color of bearing arrow	18. Vertical speed indicator
9. Wind speed and direction relative to airplane	19. Ground speed (or TTG or ET)
10. Current heading	

Figure 5. PFD from a Cessna Citation Bravo with major components labeled

Figure 6. PFD from a Boeing 737 (source: https://commons.wikimedia.org/wiki/File: Primary_Flight_Display_of_a_Boeing_737-800.png, Aug 2017)

In the context of SA, (Biella, 2009) used eye movement patterns from simulator flight scenarios with experienced and inexperienced (student) airline pilots to back reference to SA, based on the assumption that eye movements are indicators of the initial perception phase of SA. In a cockpit simulator laid out in the style of an Airbus, pilots flew eight scenarios that involved the approach, landing, and taxi phases of flights. Eye movements were measured with a head-borne iViewX system (see Figure 1) and the data were condensed into cumulative dwell times for each flight and taxi segments in nine pre-defined AOI's for different displays, charts and controls inside the cockpit and one OTW AOI. They found differences in scan patterns (dwell times on AOI's) during the separate phases of flight and several differences between student and experienced pilots. Their results are shown in Figure 7, which also provides an example of how data of this type can be presented. Introduction of an automated taxi assist system (four scenarios with the system) altered scan patters in both pilot groups but did not result in attention capture; a positive finding with respect to SA (data not shown).

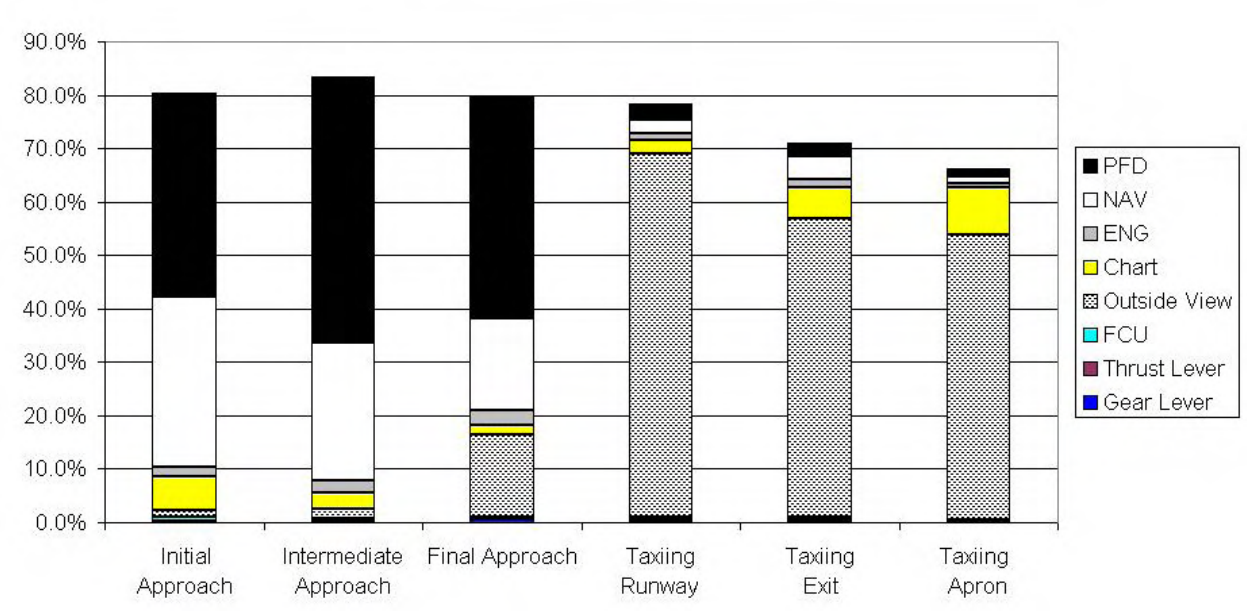


Figure 7. Results from (Biella, 2009) showing dwell times in percentage of total dwell time on the AOI's during different phases of a landing and taxi task (averages of 49 pilots)

Haslbeck, Schubert, Gontar, and Bengler (2012) looked at the relationship between pilot's manual flying skills and their eye movements during separate phases of a challenging approach and landing scenario. The principal focus of the eye movements was on the PFD, which was separated into five AOI's: attitude indicator, airspeed indicator, altitude, heading, and flight mode annunciator (Figure 8). Two groups of pilots participated, one with low levels and the other with high levels of training and practice. Eye movements were measured with a DIKALBIS eye tracker system. Glance durations (presumably fixation durations) were counted for all pilots and normalized to 100% and the median determined. Chi-squared analysis revealed differences in gaze allocation during the four phases of the flight and differences between the two pilot groups. Situation awareness was parsed into flight management annunciator (FMA), wind and speed awareness and from the eye movement data it was determined if mandatory checks were made to the AOI's that contained the necessary information during specific phases of flight (e.g. speed check between 100 feet and touchdown). The high level training and practice group completed significantly more mandatory checks suggesting they had better SA.

Yu et al. (2014) and Yu et al. (2016) used pilot scan patterns to assess the distribution of attention during air-to ground and air-to-air combat flight scenarios. In the first study, military pilots completed an air-to-ground attack mission in a fighter simulator. For data analysis, the task was divided into three phases: preparation, aiming, and weapons release/break-away. During the highest workload phase (from roll-out onto the target to break away) an unexpected event was introduced - a warning light was activated. If pilots responded to it by pressing the master caution light, their SA was judged as good, if they not, it was judged as poor. Eye movements and pupil size were recorded with a head-borne ASL 4000 (Applied Sciences Laboratory) eye tracker. Distribution of attention was assessed as number of fixation points divided into five pre-defined AOI's (four inside and one outside the cockpit. A fixation point into an area was defined as three

successive gaze points (presumably fixations) longer than 200 ms concentrated in a small region (10 × 10 pixel) of an AOI. The eye tracker also provided a measure of pupil size. Immediately after the task, pilots completed a self-assessment of perceived workload during separate phases of the mission. Differences were found in the percentage of fixations on the AOI's overall as well as differences between the separate phases of the flight. Pilots spent the most time looking at the heads-up display (HUD) followed by OTW. Pupil size was largest during the highest workload phase (aiming) and smallest during the lowest workload phase (preparation). Also, pilots with good SA, those responding to the warning light, showed lower perceived workload.

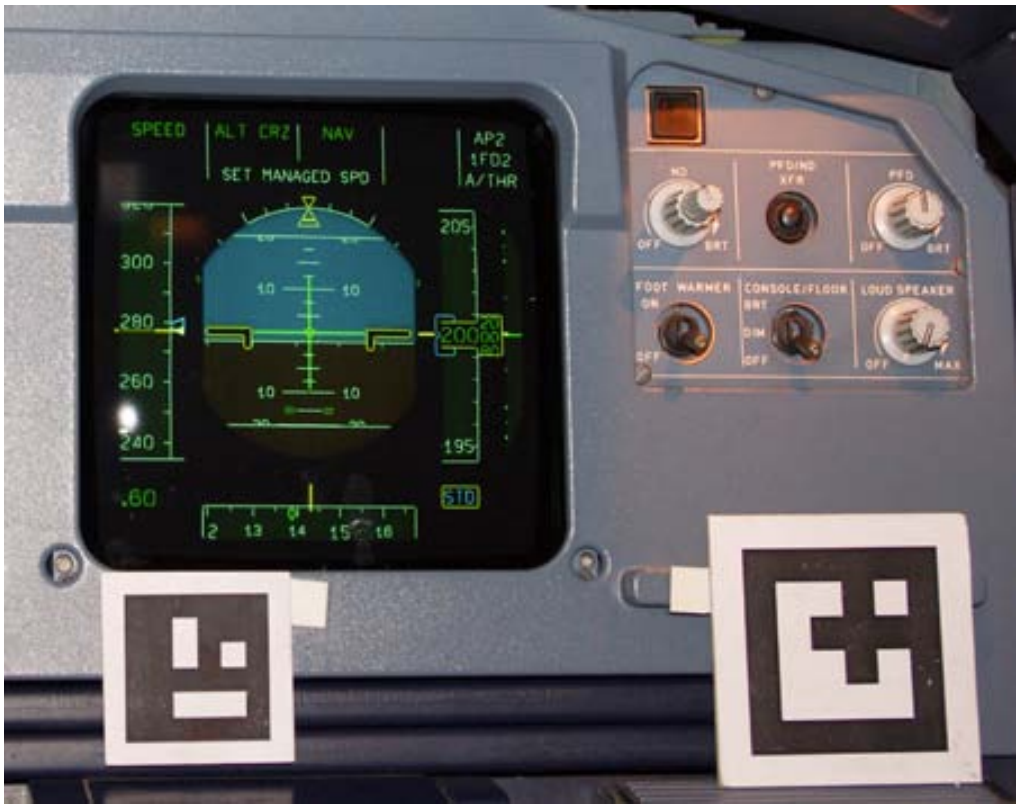


Figure 8. Photo of the primary flight display (PFD) from Haslbeck et al. (2012). The PFD was divided into five AOI's (see text for details). Also shown at the bottom are the calibration markers for the Dikalbis eye tracker

The Yu et al. (2016) study incorporated many of the same measures as Yu et al. (2014) but for an air-to-air task that involved searching for, pursuing and locking on to a moving airborne target while performing air-to-air flight tasks. As in the earlier study, differences were found in the percentage of fixations on the AOI's overall and in addition, fixation durations. Found were also differences in fixations on the AOIs and fixation durations between the three phases of flight with the shortest fixation durations during target lock on activity. Different trends in fixation were found for novice and experienced pilots. Experienced pilots focused more on the HUD to extract information about the target while novice pilots spent more time looking OTW, although overall both groups more time was spent OTW than on the HUD. Pupil size differed between phases with the largest sizes being during lock-on. SA differed between the groups with 77% of experienced

pilots responding to the unexpected warning indicator compared to 23% of novice pilots. Whether the target was hit or not was recorded but those results were not reported.

One difference between the two Yu studies (Yu et al., 2014; Yu et al., 2016) is that pilots spent more time looking out the window during the air-to-air task while the HUD was the major focus of attention in the air-to-ground task. The difference in gaze allocation was attributed to differences in the operational context. The task of search and pursuit of a moving aerial target was best facilitated through OTW viewing as opposed to search and weapons lock-on to a stationary ground target, which was facilitated best by using the HUD. The differences in gaze patterns between the two flight scenarios in these two studies emphasize the point that it is important to take into consideration not only the phase of flight but also the operational context. As another example of operational context, Yang, Kennedy, Sullivan, and Fricker (2013) had pilots fly overland routes with difficult and easy sections and found evidence that more experienced pilots changed their gaze pattern during the difficult section to more out-the-window viewing and less time on the map.

The goal of studies by Diez et al. (2001) and Sarter, Mumaw, and Wickens (2007) was to gather information about the interaction of pilots with automation for developing models of pilot cognition with automation. The models would then be used to evaluate interventions designed to reduce automation problems. Eye movements were used as the metric to assess what information pilots were attending to, since this is the first phase in the cognitive cycle (also the first, perceptual, phase in the models of SA). Diez et al. (2001) used commercial airline pilots, who flew two scenarios in a simple desktop simulator of a Boeing 747-400. Scenario one involved take-off, climb, cruise, descend, and approach phases of a flight. Scenario two, only involved the descent and approach phases of flight. Eye movements were recorded with an ASL 504 eye tracker and the mean fixation duration on AOI's extracted. In addition, a "freeze" technique, commonly used in studies of SA, was employed with pilots interrupted six times during a scenario and asked to recall details about the flight situation. Sarter et al. (2007), extended the Diez et al. (2001) study by employing a more sophisticated simulator and flight scenarios that involved twelve challenging automation flight-related events (e.g., a loss of glide slope and glide slope diamond). Eye movements were recorded with an ASL 400 eye tracker and data extracted were total dwell times in five AOI's as well as percentage dwell times during separate phases of flight. Both studies found the PFD was the most attended to instrument but also that scan patterns differed for the separate phases of flight. For example, pilots spent little time looking OTW except during the final approach phase.

Lefrancois, Matton, Gourinat, Peysakhovich, and Causse (2016) were also interested in the interaction of airline pilots with automation, but from the perspective that over-reliance and use has resulted in the loss of manual flying skills. They used eye-tracking data to assess gaze distribution during manual approaches to test a hypothesis that gaze patterns would differ between pilots who flew a stabilized vs unstabilized approach (made a missed-approach go-around). Eye movements were recorded with a Pertech system and gaze allocation (% time) for eight AOI's was determined. Commercial airline pilots flew an approach and landing in a full motion simulator under instrument landing system (ILS) conditions. A Precision Approach Path Indicator (PAPI) was present on the left side of the runway to provide feedback about the terminal approach phase. Airspeed and altitude at runway threshold and touchdown points were extracted from the simulator data. Four of the twenty pilots flew an unstable approach and their results were compared to the

four pilots who flew the best stable approach. Between the groups, differences in airspeed, height above runway threshold, and touchdown point were found, as well as differences in gaze allocation. One of the largest differences between the two groups was the standard deviations of the gaze distribution percentages. The four best pilots had very similar gaze distributions, while the worst pilots did not follow a consistent pattern regarding where they looked during this tasking (Figure 9).

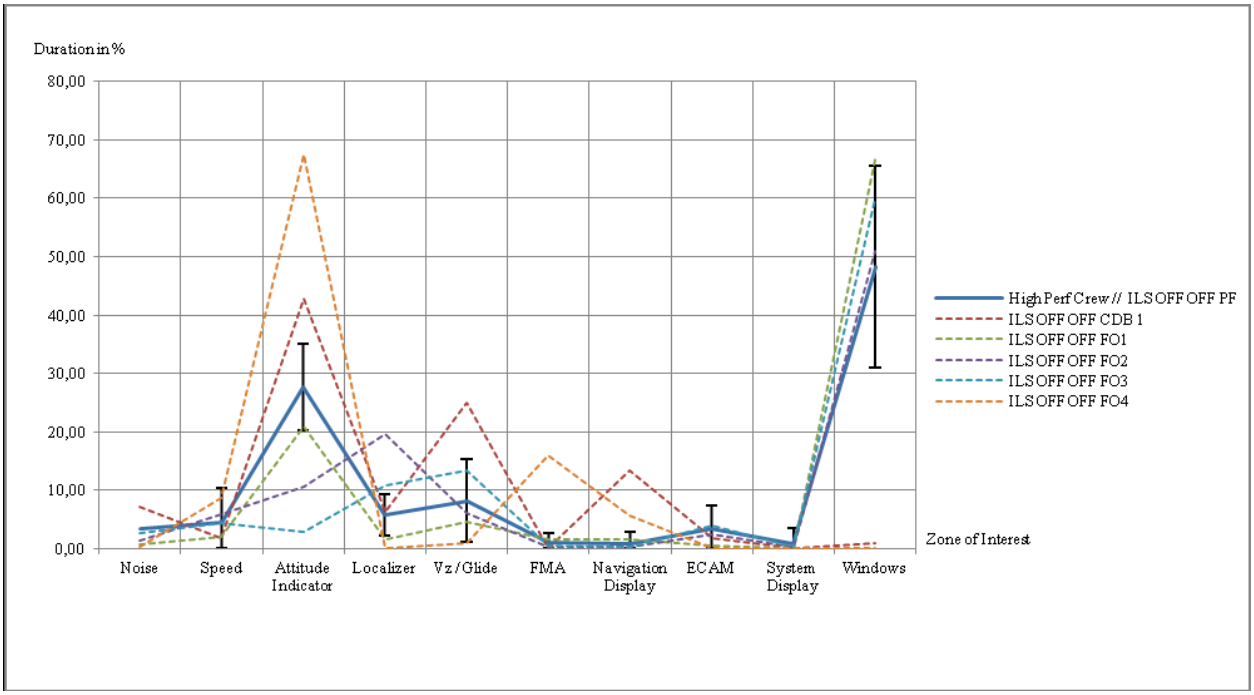


Figure 9. Percent of time spent looking (dwell time) at the AOI's defined in the Lefrancois et al. (2016). The solid blue line is the average of the four best pilots. The dotted lines are individual data for the four pilots who flew unstable approaches

4.2.2 Training strategies.

Multiple studies analyzed eye movement patterns as well as flight performance in pilots of varying levels of expertise (Bellenkes, Wickens, & Kramer, 1997; Chuang, Nieuwenhuizen, & Bülthoff, 2013; Kasarskis, Stehwien, Hickox, Aretz, & Wickens, 2001; Kirby, Kennedy, & Yang, 2014; Sullivan, Yang, Day, & Kennedy, 2011; Yang et al., 2013). The goals of these studies were to determine the relationship between flight control and instrument scanning behavior for the purpose of developing training strategies for instrument monitoring, and to inform changes in instrument layout. Two studies were conducted in fixed wing simulators (Bellenkes et al., 1997; Kasarskis et al., 2001) and the others were in rotary wing simulators. Although, different flight scenarios were tested, common to all were maneuvers to change altitude, speed, and heading. All the studies found differences between novice and expert pilots for various aspects of scanning, including scan patterns, scan frequencies and dwell times. A general finding was that expert pilots made more frequent fixations but had shorter dwell times than novices. Flight performance was found to be better in experts in some studies (Bellenkes et al., 1997; Kasarskis et al., 2001; Kirby et al., 2014), but not others (Sullivan et al., 2011; Yang et al., 2013); a difference likely related to the level of sophistication of the simulator flight controls.

The flight task in the Bellenkes et al. (1997) study was a seven segment flight that involved various combinations of straight and level constant speed segments and maneuver segments involving heading, altitude and/or airspeed changes. In the Kasarskis et al. (2001) study, the task was a turn into final approach and landing under VFR conditions that was repeated 12 times. Both studies employed desktop simulators with simple flight controls, such as a flight yoke to control pitch, roll and airspeed and both used ASL eye trackers to record eye movements and the scene. The flight performance data collected in both studies ware similar in that lateral and longitudinal deviations from a flight path were used as indicators of flight performance. Bellenkes et al. (1997) used root mean squared error (RMSE) for altitude, heading and airspeed to track performance on three axes, vertical, lateral and longitudinal, respectively. Kasarskis et al. (2001) used a weighted formula to score landing performance based on lateral and longitudinal deviations relative to the width of the runway and length of the landing zone. Figure 10 shows the results for the expert pilots and illustrates a method for presenting this type of data.

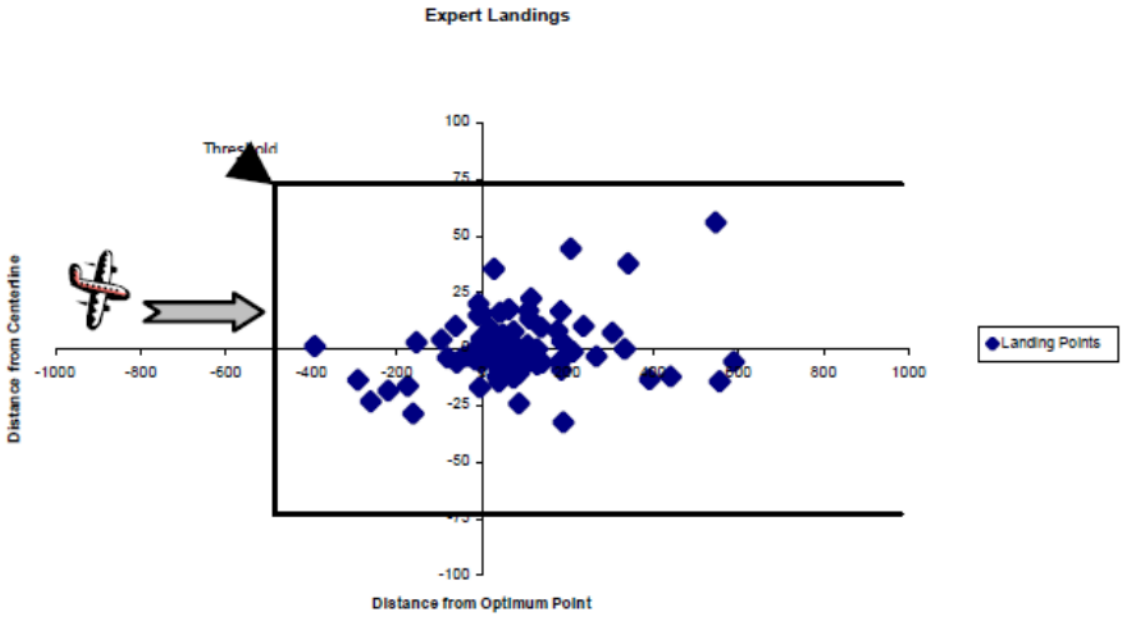


Figure 10. Landing results for expert pilots taken from Kasarskis et al. (2001). The optimal touchdown point is at the intersection of the axes, which indicates distance along the runway (x-axis) and distance from centerline (y-axis). The start of the runway threshold is the vertical line indicated by the filled triangle and the horizontal lines show the runway width.

The flight tasks in the rotary wing simulators included a low level over-land navigation task at moderate speed but with ambiguous terrain (Sullivan et al., 2011; Yang et al., 2013), a low level, high speed over-land navigation task with varied terrain (Kirby, Kennedy, & Yang, 2013), and a simulated commuter flight from a suburban airport to a city (Chuang et al., 2013). An example of an over-land navigation task is shown in Figure 11. Simulator sophistication varied from relatively simple with control provided by a side mounted joystick (Sullivan et al., 2011; Yang et al., 2013)

to simulators with full controls, cyclic stick, collective lever and foot pedals (Chuang et al., 2013; Kirby et al., 2013). Flight performance outcome measures were RMSE for flight trajectory and flight duration error calculated as actual trajectory/duration minus ideal (Sullivan et al., 2011; Yang et al., 2013), standard deviation of altitude (Kirby et al., 2013), and RMSE between desired and actual altitude and airspeed (Chuang et al., 2013). All the studies used FaceLAB (Seeing Machines) eye trackers.

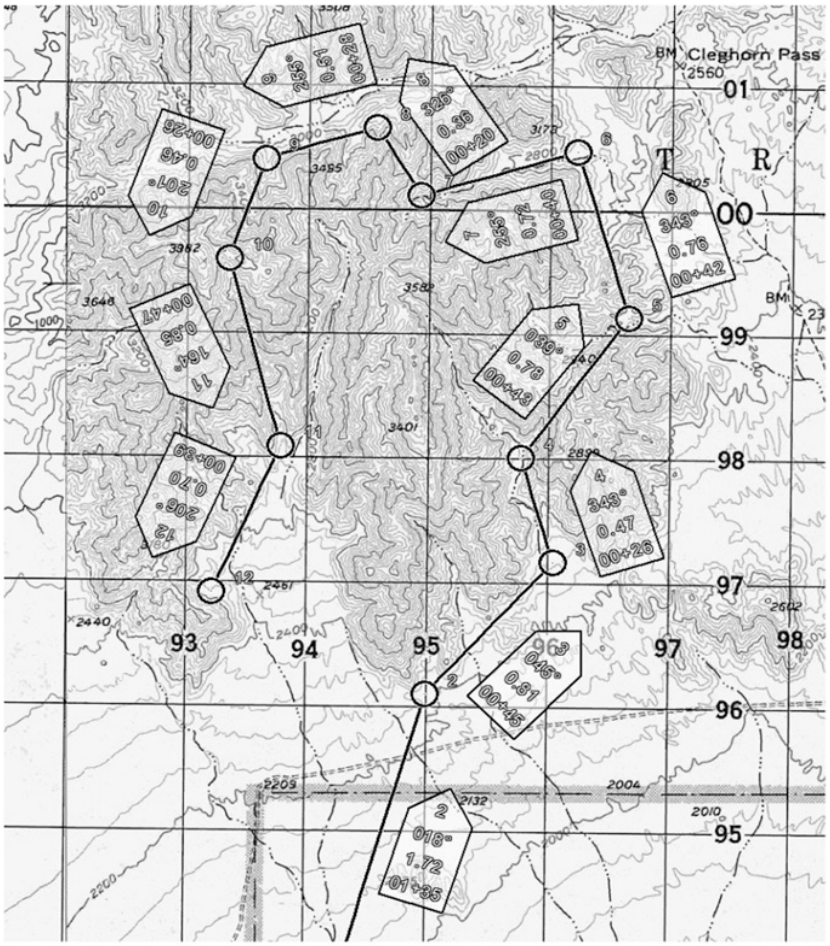


Figure 11. Overland navigation task from Sullivan et al. (2011). The "doghouse" between waypoint pairs indicates from top to bottom the next waypoint number, recommended heading to next waypoint, distance and estimated time of flight based on a fixed speed

4.2.3 Displays.

Another application of eye trackers in the flight simulation environment has been in studies of the effects of display characteristics on information acquisition. The studies reviewed investigated the effects of clutter on a persons or pilot's ability to extract information from a display. An example of differences in clutter at three levels in the PFD is illustrated in Figure 12.

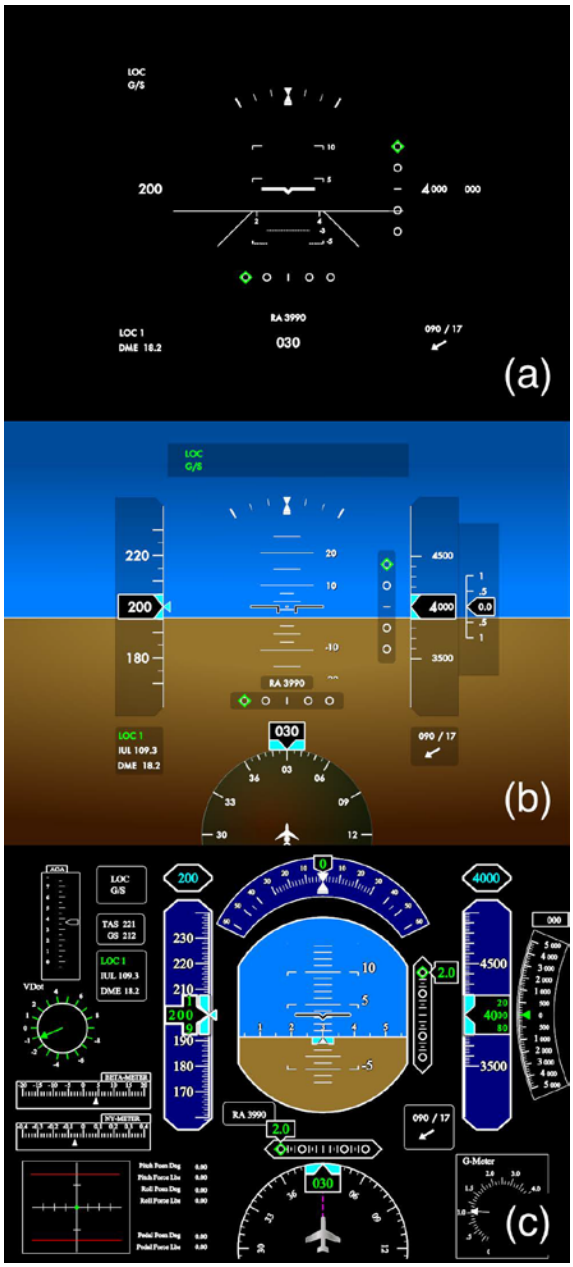


Figure 12. Example of three levels of clutter in the primary flight display in (a) low, (b) medium (c) high clutter levels (taken from Doyon-Poulin, Ouellette, & Robert, 2014)

The purpose of the Moacdieh and Sarter (2012) study was to determine which metrics, from a list extracted from the literature, were sensitive to clutter effects and might be used in future display evaluations. The subjects in their first study were not pilots, but engineering students whose task was to find Waldo in a set of images taken from the book "*Where's Waldo*" (Handford, 1997a, 1997b). The images had two levels of clutter that were either static or dynamic. In the dynamic images, parts of the background were animated to move in translational or rotational motion while parts of the display remained static, including where the target Waldo was located. Subjects

responded with a key press when they had found the target and then indicated its location with a mouse. Response time and error rates were determined. In addition, an ASL eye tracker was used to record eye movements (Figure 13). The images were divided into nine AOI's (3 × 3 grid) for most of the analysis but a finer 9 × 13 grid was also used. The target AOI contained the Waldo image. The results indicated that response times and error rates were higher in the high clutter condition, while there was no effect of the dynamic components on performance. They also calculated the numerous eye movement metrics used by other investigators and determined if they were significantly affected by clutter, which the majority were. The conclusion was that a variety of eye movement metrics could be used to assess the effects of clutter in displays.

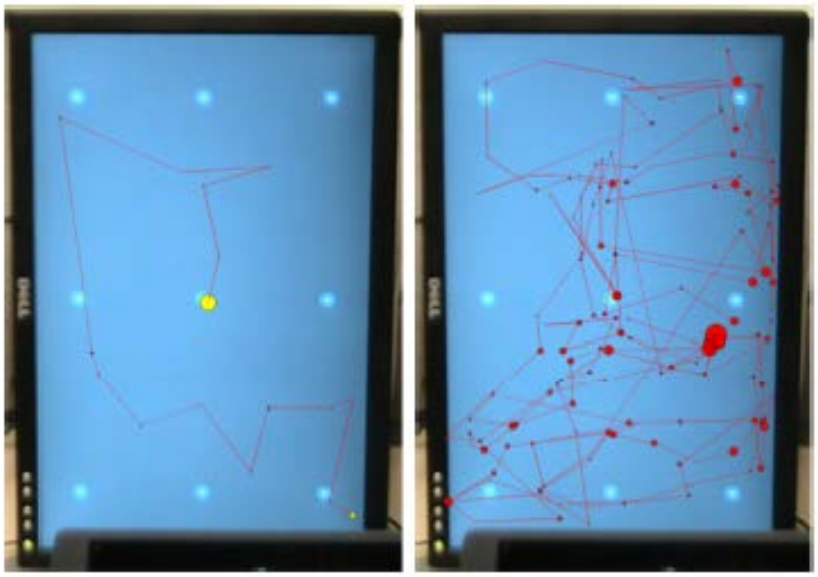


Figure 13. Example of differences in number of eye movements (red lines between fixation dots) made by a subject in the low clutter (left) and high clutter (right) conditions in the find Waldo task. Circle size relative to fixation dwell time. (Moacdieh & Sarter, 2012)

In a follow-on study with pilots, Moacdieh, Prinet, and Sarter (2013) investigated clutter using a combination of eye movement data, simulator performance, and self-reporting. Three levels of clutter in the PFD were tested using three groups of instrument rated pilots. The flying task involved take-off, cruise, altitude change, and landing and involved low and high workload conditions. Low workload was the cruise phase and high workload was associated with take-off, landing, altitude change, and high turbulence. Twenty-two visual alerts (no auditory component) were presented during the flight, equally divided between the workload conditions and pilots had to acknowledge them as quickly as possible by pressing a button on the yoke. The alerts appeared in different areas of the PFD and included notifications, instrument failures (red X over the instrument), altitude alerts (altitude changes color), and yellow or red alerts in different regions of the PFD. An ASL eye tracker was used to record eye movements. Response time, percent missed alerts, and three eye movement parameters were the dependent variables. One eye movement metric was spatial density, determined by dividing the PFD into an array of cells, with each cell

being 17 × 17 pixels and then dividing the number of cells with at least one fixation by the total number of cells. A large density meant a substantial percentage of cells were scanned. The other metrics were percent fixation duration on the FMA and number of transitions between pre-defined AOI's on the PFD, such as the airspeed indicator, flight director, and altitude indicator. Increasing clutter increased response times to alerts as did increased workload and more alerts were missed in the high workload condition. These investigators found a significant effect of clutter on spatial density such that it increased (more scanning) with increasing clutter. The number of transitions also increased in the high workload condition. Pilots rated clutter along six dimensions (e.g., redundancy, colorfulness) and overall felt the medium level clutter provided the most information. Data on flight parameters across the three clutter conditions was not reported.

Doyon-Poulin et al. (2014) took a similar approach by using PFD displays with three levels of clutter. (Figure 12). They had pilots fly manual approaches to a major airport and to increase workload and make them use the PFD to obtain guidance cues (glide slope, localizer), most automation was turned off. The independent variable was level of clutter of the PFD. Dependent variables were pilot assessments of clutter and workload as well as deviation from the flight path recorded by the simulator. A FaceLAB eye tracker recorded scan patterns on the PFD screen. Flight parameters were localizer deviation, glide slope deviation, stick activity, and speed control. Eye movement parameters were fixation points on the screen for the different PFD's, mean difference in time between fixations on the screen, cumulative fixation points, mean fixation time, and transitions between zones of the PFD. They also defined a small AOI for the localizer (part of the PFD) and analyzed cumulative fixation time on it for the three clutter levels. Clutter ratings were consistent with the physical amount of clutter on the PFDs and pilots considered the medium clutter display to be most pleasing and to require the least workload to extract information. Localizer error was less when the medium clutter PFD was used and cumulative fixation times highest. No other parameters were statistically significant in relation to clutter.

Ahlstrom and Dworsky (2012) used a Micro-Jet simulator configured as a Cessna 172 with glass cockpit displays to assess three different formats for presenting meteorological information (number of colors, color format, types of symbology) with respect to workload and flight parameters. Twenty-five instrument rated general aviation pilots were assigned randomly to one of the weather display format groups. Pilots twice flew an overland route beginning mid-flight (no takeoff or landing) and in VMC and after five minutes into IMC for 25 minutes with the flight ending back in VMC. Data extracted from the simulator included course deviations in latitude/longitude from a straight flight path (weather avoidance), mean distance to a specified rain level, frequency of usage of the zoom feature on the weather displays, aircraft position, mean altitude, and mean heading.

Eye movement parameters were also recorded with an ASL mobile eye tracker equipped with a head tracker. With respect to eye movements, the investigators concluded, based on previous work that eye movement parameters can serve as proxies for cognitive and visual workload. To support this they noted that Ahlstrom & Friedman-Berg, 2006 found that as cognitive workload increases, blink rate, blink duration, and saccade duration decrease. In contrast, pupil diameter, the number of saccades, and the frequency of long fixations all increase. The eye movement parameters recorded in the study by Ahlstrom & Dworsky, 2012 included number and duration of fixations, number of saccades and saccade distances, pupil diameter and eye movement workload (number of points of gaze). Areas of interest were also defined for OTW, the glass display, weather display,

and cockpit console and the same fixation and saccade parameters listed above extracted for each. They also assessed physiological workload using an fNIR system to capture blood oxygenation changes.

The flight was divided into three legs for analysis. A Bayesian model comparison analysis was conducted which allowed a straightforward comparison between displays, but uniquely allows evidence in support of the null hypothesis or the alternative to be stated in the form of an odds ratio (null/alternative). A complete discussion of the analysis and results is beyond the scope of this paper but the main conclusion was that differences in flight deviations, scan patterns, and cognitive workload were found between the three display formats. With respect to eye movements they found one display required more saccades and fixations than the other two and that subjects in that display group showed a different scan pattern that focused on glass and weather display AOI's, suggesting that this weather information format required more cognitive effort to extract information. Also reported was a problem with the eye tracker in that it could not be used with subjects wearing glasses and this limited data collection to 15 out of the 25 subjects.

4.3 Simulator Flight Performance Metrics

Studies in this group covered a range of topics from evaluation of display characteristics to effects of degraded biological conditions, with simulator flight performance being the primary or one of the major outcome measures.

4.3.1 Display evaluation.

Numerous studies assessing the effects of different display characteristics such as color coding, clutter, and letter and symbol size have been conducted. Several studies on clutter were reviewed earlier in the eye movement section. In general, display evaluations in the aviation domain have been in the context of effects on SA. (Salud, 2013) provides a meta-analysis of different display characteristics and their effects on SA as well as a review of the literature used to develop her libraries. Although the purpose of her thesis was to provide next generation display designers with a tool for estimating the effects of different display elements and presentation modalities on SA, the discussion of specific display properties and elements and how effects were evaluated may be potentially useful in developing display simulations for LEP evaluation. For example, highlighting a single target with a unique color or high intensity compared to no-highlighting results in an 84% decrease in detection time. Would a similar decrease be found with LEP or would the highlighting effect be reduced? For comparison with the clutter evaluations, a sample of studies that examined the effects of color-coding and one on presentation format in displays were reviewed.

Christ (1975) reviewed the early display color-coding literature from the 1950's to the 1970's and concluded that the consensus was that color-coding greatly enhanced search speeds even if it was a redundant cue. However, for identification tasks, color-coding increased speed when it was unique, but when it was redundant with other cues its value was questionable. This problem was taken up a decade later by Luder and Barber (1984) who used a dual task paradigm, where subjects performed a compensatory tracking task while also performing search and identification tasks, to judge fuel status on either a monochrome or a color systems management display. Color-coding, even when it was redundant with other cues (shape), improved search performance but yielded no benefit on the identification task. In fact, in some cases color-coding interfered with identification because it was the more salient cue in a situation that required identification to be based on a non-color cue. One other benefit of color-coding was that the color group performed much better on

the tracking task, presumably because the color-coding of the fuel display allowed more cognitive resources to be devoted to the tracking task. In terms of LEP effects evaluation, the task used by Luder and Barber (1984) in modified form might be considered as a task that can be performed by non-aviators and will be discussed in more detail later.

Another study that looked at color-coding was that of Post, Geiselman, and Goodyear (1999) who looked at the advantage of color-coding weapons symbology for helmet-mounted displays (HMD). The flight scenario was a multi-aircraft air-to-air engagement with the pilot-subjects flying in an F-16 cockpit mock up. The color-coding scheme for the HMD is described in detail but the primary interest was whether color-coding the target designator box so that red means shoot provided any advantage over the appearance of a monochrome green shoot cue that was the same color as all other symbology and text on the display. Overall, the red means shoot color coding resulted in significantly faster shots against both fighter and bomber aircraft as well as longer missile release distances. There were some differences between bomber and fighter targets depending on the status of the red shoot cue (steady or flashing at 5 Hz), but these were attributed to a difference in tactics for the two roles. Other colors used in the code were green and yellow, with green indicating targets were not in range and alternating yellow and green indicating they were in range but at the limits of the missile range and maneuvering envelope.

Olmos, Wickens, and Chudy (2000) used a simple simulator to compare a conventional two-dimensional coplanar tactical display to two, three-dimensional perspective displays. The simulator consisted of a Silicon Graphics display and a two-degree-of-freedom (2df) joystick to control navigation of an airplane with the dynamics of a light aircraft. The most relevant aspects of this study to LEP evaluation is the use of a simple simulator, the flight task and to a lesser extent the information displayed. The 2df joystick allowed subjects to climb, descend, and bank the aircraft up to 90°. Bank angle was coupled with pitch so that the downward pitch of the nose was appropriate for the bank angle. The aircraft was held at constant power but speed could vary up or down as it naturally would during descent and climb, respectively. There were no rudder or throttle controls. Small altitude and airspeed indicators were added to the top of the tactical display.

These simple controls allowed subjects to navigate the plane through eight waypoints. Waypoints appeared sequentially as flashing yellow cubes on the screen and varied in location and altitude relative to the plane position. The subject's task was to intercept a waypoint, whereupon the next waypoint appeared. Two of the legs of the flight were straight but the other six required maneuvering the aircraft around permanent obstacles. In addition, on two of the maneuver legs an external threat (another aircraft) suddenly appeared and had to be avoided. The subject also had to call out the relative altitude (high, level, low) and heading (toward or away) of the threat. On two other of these legs, a pop-up conflict in the form of a hazard zone (radar cone) suddenly appeared and had to be avoided. Performance measures were total time within each leg, contacts with surrounding terrain or hazards and duration of contact, external threat response time and accuracy, and XYZ position of the aircraft, which was used to assess quality of maneuvers (climb, turn) to navigate around hazards relative to an ideal maneuver. The result of the first experiment indicated deficiencies in all the displays, and these were corrected. A second experiment was performed, and these data verified that the corrections had solved the problems.

4.3.2 Degraded biological conditions.

Over the years there have been a large number of studies that have looked at the relationships between flight performance, attention, and cognitive function under degraded biological conditions such as sleep deprivation (Caldwell, Caldwell, Brown, & Smith, 2004; Lopez, Previc, Fischer, Heitz, & Engle, 2012; Russo et al., 2004), drug and alcohol consumption (Mumenthaler et al., 2003; Yesavage et al., 2002) and dehydration (Lindseth, Lindseth, Petros, Jensen, & Caspers, 2013). Since the focus of these studies was on factors not likely to be included in the evaluation of LEP, only a small selection were reviewed as examples.

Caldwell et al. (2004) and Russo et al. (2004) both investigated the effects of continuous wakefulness on cognitive function and flight performance. These study data were in good agreement showing that significant decrements in both domains begin to occur after approximately 20 hours of continuous wakefulness. The flight parameters evaluated in the Caldwell et al. (2004) study were altitude, airspeed, vertical velocity, heading and roll during various turning maneuvers, climbs and level flight relative to ideal flight paths. Russo et al., measured azimuth deviations during an aerial re-fueling maneuver. Of potentially more relevance to LEP issues are the visually-mediated tests that were performed to provide an indicator of cognitive function. In the Caldwell et al. (2004) study, an unstable compensatory tracking task from the Multi-Attribute Task Battery (Comstock & Arnegard, 1992) was performed during which subjects concurrently monitored warning lights and dials and responded to various auditory requests to adjust radio frequency and perform fuel transfer tasks. Data collected were tracking errors, response times, time-outs, false alarms, and accuracy rates. Russo et al. (2004) used what they called a choice visual perceptual task where pilots had to respond to single and multiple flashes of light superimposed on the instrument panel while performing the high cognitive load re-fueling maneuver. The light stimuli were presented randomly from 0 to 75° along the horizontal meridian against the dark background of the display. Response time and omissions were recorded, although only the latter was used to assess visual neglect.

4.3.3 Age, expertise/experience, and practice.

Aging is another biological condition that results in declines in cognitive function, visual function, and motor function, all of which are closely associated with flying aircraft. Tolton (2014) provides a helpful review of the literature on the effects of age, expertise and cognitive function on flight performance. Most simulator-based studies agree that as age increases, flight performance decreases (Taylor, Kennedy, Noda, & Yesavage, 2007; Tolton, 2014; Yesavage, Taylor, Mumenthaler, Noda, & O'Hara, 1999). However, the amount of performance degradation that age can statically account for has generally been found to be small (e.g., < 25% reported by Yesavage et al., 1999), suggesting other factors such as cognitive and motor function are also involved (Taylor et al., 2007). Since older pilots generally have more experience, several studies investigated if expertise could compensate for age in simulator performance. While most studies found that experience and performance were positively associated, there was not a consensus that long-term experience could compensate for degraded performance from advancing age.

Yesavage et al. (1999) tested the hypothesis that increased age is associated with decreased performance on flight simulator tasks. In a cross-sectional study design, the performance of 100 pilots aged 50-69 was investigated using a simulated flight task that consisted of take-off, cruise, approach and landing. During the cruise segment, sixteen en-route course changes were required and three emergency situations occurred (e.g. sudden appearance of air traffic). The simulator had

the instruments, controls (yoke, rudder pedals, and throttle) of a small, single-engine aircraft like a Cessna 172. Twenty-three flight performance variables (errors or deviations from ideal or assigned values; e.g., altitude, heading, airspeed, reaction time) were recorded by the simulator. These were converted to z-scores and aggregated into eight flight component scores representing takeoff, course deviation, communication frequency, traffic avoidance, cockpit monitoring, approach corrections, runway alignment, and rate of decent at touchdown. Five of the scores were summed and the mean used as a summary score of overall performance. Takeoff and landing elements were not included in the summary score because they had previously been shown not to have good test-retest reliability. Although performance declined with age and the results supported their hypothesis, age accounted for only 22% of the variance in performance, suggesting other factors are involved in determining flight performance.

In a companion study with the same subject sample, Taylor, O'Hara, Mumenthaler, and Yesavage (2000) found significant relationships between several components of the CogScreen Aeromedical Edition (AE) and flight performance. The CogScreen AE is a battery of tests that assess perceptual, cognitive and information processing abilities. Four CogScreen variables accounted for 45% of the variance in the same flight summary score described by (Yesavage et al., 1999); speed/working memory, visual associative memory, motor coordination and tracking. All were significantly associated with age and adding age to the multi-factor model improved predictions, although the specific amount of improvement was not stated.

Taylor et al. (2007) and Tolton (2014) addressed whether expertise can compensate for age-related declines in simulator performance. Using the same simulator and flight performance metrics described in their earlier studies (Taylor et al., 2000; Yesavage et al., 1999), Taylor et al. (2007)partitioned 118 pilots, ages 40-69, into three expertise groups, and each pilot completed a simulated flight every year over a three year period. Subjects also completed a cognitive assessment (CogScreen-AE) and tests of information processing speed. Tolton (2014) used a similar methodology and both studies found, as expected, that older pilots performed worse than younger pilots on the cognitive, information processing speed, and flight tests. They also found that performance of older pilots declined less over time than younger ones. In addition, they found that expert pilots performed better at baseline than other pilots and showed less decline over time. However, the results did not allow them to conclude that expertise moderated the effects of age on aviation performance.

Haslbeck, Kirchner, Schubert, and Bengler (2014) did not study age per se, but instead looked at the effects of practice and training on manual flying skills of airline pilots. In an earlier paper using the same subject sample they looked for differences in scan patterns between pilot groups with high or low levels of practice and training (Halsbeck et al., 2012). The 2014 study focused on manual flying performance with the dependent measure being deviation from the ideal flight path, which was determined from glide slope (vertical guidance) and localizer (lateral guidance) deviations. Both raw deviation and RMSE data were analyzed and revealed that the high practice and training level group (First Officers) performed significantly better than the low practice, low training group (Captains) even though the Captains had more overall experience (Figure 14).

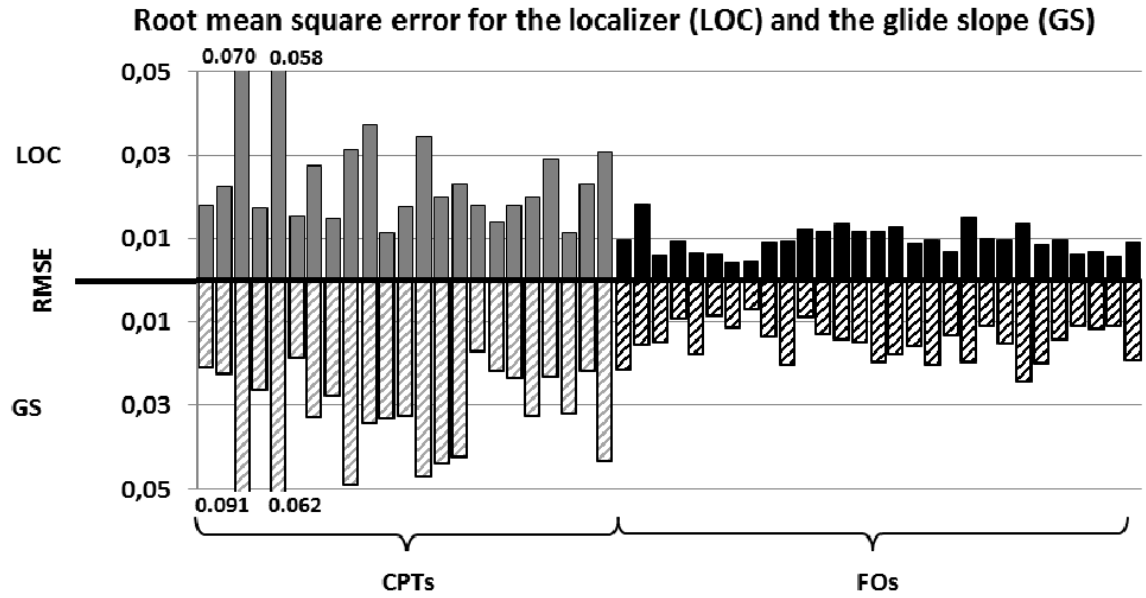


Figure 14. Summary of Haslbeck et al. (2014) results showing deviations from the localizer and glide slope in RMSE for the two pilot groups

4.3.4 What are the best metrics?

A key question regarding the use simulators is what simulator metrics are the best to use for assessing performance? Several studies that were reviewed addressed this question and the results depended on both the flight task itself, and the type of aircraft being simulated, e.g. fixed or rotary wing.

Crognale and Krebs (2008) had helicopter pilots flying a FlyIt simulator complete flight scenarios that involved flying from VMC into IMC conditions to determine how weather changes effect pilot performance. The main finding was, somewhat unsurprisingly, that reduction in visibility resulted in declines in pilot performance. Of note is their data analysis which involved an interesting approach, it used two types of data analysis, a "power" analysis based on power in a Fourier transform of the data and an error analysis that compared error rate between the IMC and VMC conditions. The inverse of the Fourier power in the spectrum was taken as an objective measure of stability or aircraft control performance and it can be applied to both output and control input data. It also has the advantage or providing a continuous measure over time and not just a measure when some criterion is exceeded. Eight measures from the power and error analysis were chosen for analysis: pitch power, pitch error rate, fore/aft cyclic power, bank power, bank error rate, lateral cyclic movement power, pedal movement power, and vertical airspeed error rate. In several cases they found that the error analysis did not reveal a significant effect of IMC conditions because criterion control levels were not exceeded, but the power analysis revealed that significantly increased effort was required in IMC for the pilot to maintain good control of the aircraft.

Lee (2010) had instructor and student pilots fly the approach and landing phases of flight in a Cessna 172 stimulator over two different approach areas (unpopulated vs. populated) at two different approach angles (normal = 3° and steep = 4.5°). Pilots were instructed to fly a straight-

in approach, and land 1,000 feet from the runway threshold at a speed of 65 mph. PAPI lights were used to provide vertical guidance information and cockpit localizer indicators provided horizontal guidance information. Heart rate (Polar S810i) was used to measure stress level and the NASA-task load index, scored on six-sub dimensions described in Moroney, Biers, and Eggemeier (1995), were used to measure subjective workload. Data extracted from the simulator were landing distance and speed (landing performance), and above glide path tracking performance. Flying over populated areas resulted in poorer performance, increased stress, and higher workload. Comparable results were found for flying a steep glide angle. They concluded that metrics of stress and workload were important to assess when evaluating the effects of new flight procedures on pilot performance.

Appel, Schubert, and Hutting (2012) investigated whether the manual flying skills of airline pilots performing approach and landing could be assessed from steering inputs. This study design contrasted with traditional measures of landing performance, which are largely based on glide slope deviations and instructor's use of that information to rate pilot performance. They compared steering inputs for three pilots in the three instructor rating categories (good to inferior performance). Steering inputs for roll and pitch included number of peaks, the ratio of positive to negative peaks, number of valleys and peak to valley ratio, amplitude (max, mean and variance) and the frequency of steering inputs (max, mean and variance). Figure 15 illustrates how some of the steering inputs were defined.

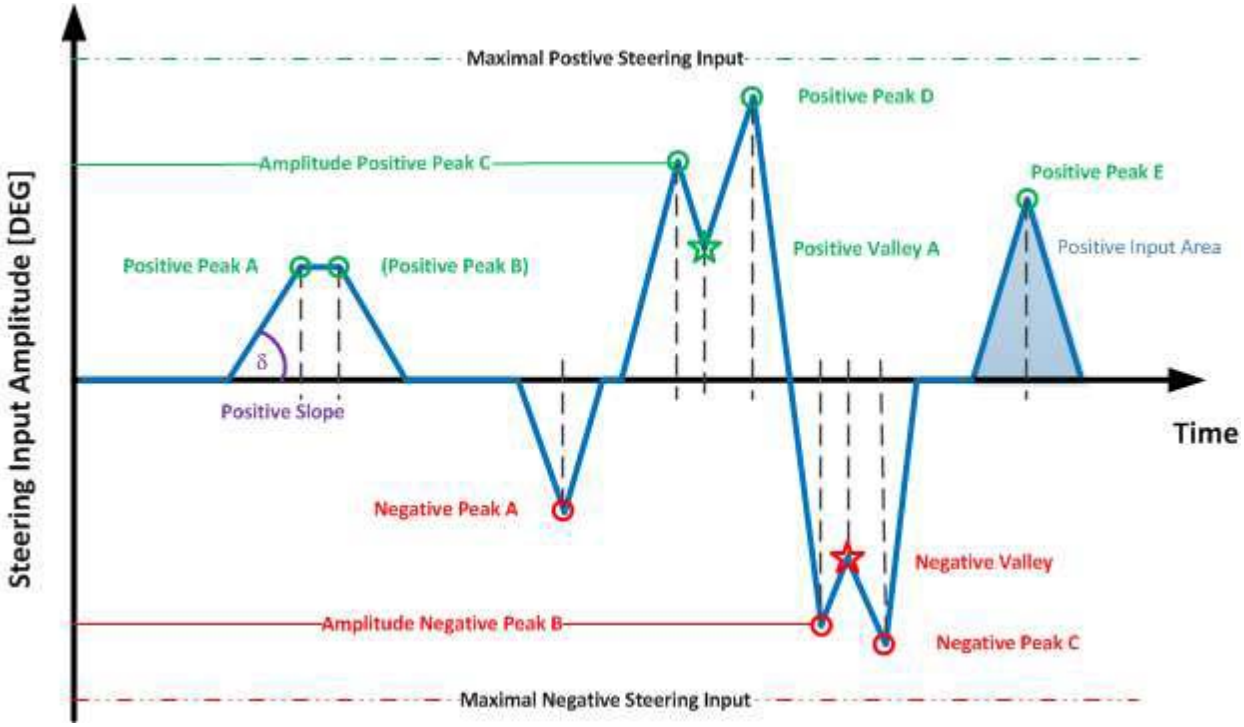


Figure 15. Example of features of steering inputs from Appel et al. (2012)

The steering amplitude and frequency data were analyzed in terms of the power spectrum for roll and pitch, and the inputs for roll and pitch in degrees were plotted in polar coordinates to provide a dual input graph. These investigators found that the number of roll steering inputs and amplitude

variance correlated with track error as did the valley to peak input ratio. Also that a number of different inputs varied with instructor ratings of the pilot's flying skills. Overall the study indicated that manual flying skills could be assessed by looking at steering inputs for roll and pitch. The use of steering input data provides something of a novel approach to assessing flying skills.

4.3.5 Training

Simulator flight metrics have also been used to assess the effectiveness different strategies applied to pilot training. For example, Khan, Rossi, Heath, Ali, and Ward (2006) looked at the effects of providing OTW visual cues during training to land an aircraft and making a level 360° turn. The OTW cues were virtual hoops presented on the screen that subjects flew through as part of the training to teach aircraft control available in the Microsoft Flight Simulator (MSFS) software. Experimental subjects were students with no prior flight experience, but screened for flying aptitude based on their performance in three straight and level flights. The subjects were divided into three groups, one group received eight sessions of landing training without OTW visual cues, another with OTW cues always present, and the third with OTW cues present on half the training sessions. For the 360° turn, one group always had cues present and the other two groups flew some flights with visual cues and others without cues. The difference between the OTW cue groups was the density of cues present. The performance measures in the Cessna simulator for the landing was the sum of RMSE in air speed, rate of decent and runway alignment and for the turn it was the RMSE from the ideal path based on bank angle and altitude loss for the 360° level turn.

Results indicated that for the landing task, subjects who trained with OTW cues in each training session performed more poorly than controls or subjects who trained with OTW cues only during some of the sessions. In contrast, for the level turn, OTW cues provided a performance advantage compared to training with no cues, but the advantage was greatest for the group that trained with a lower density of cues. However, it should be noted that for turn training the OTW cue groups did not use cues in each training session. For the landing scenario, OTW cues were hypothesized to provide a source of focus such that subjects were more concerned about flying through the hoops than monitoring the instrument panel to determine that proper flight parameters were being maintained. In the turn task, OTW cues were present only in about half of the training sessions, giving subject the opportunity to learn to monitor their instruments rather than focusing on the hoops but, at the same time, sporadic use of the OTW cues was a better strategy than training without any cues.

4.3.6 Startle and laser exposure.

Martin, Murray, and Bates (2012) investigated the effect of startle on pilot performance during a critical phase of flight. Instrument rated 737 pilots completed hand flown ILS approaches in a 737 simulator. Weather conditions were set to require a standard procedure missed approach when a decision altitude (DA) of 200 feet was reached, due the fact that the cloud base was set at 100 feet above ground level. During the first approach at 40 feet above DA a cargo fire warning bell coincident with a loud bang provided a startle stimulus. No startle stimulus was delivered on a second approach. The minimum altitude with startle was compared to that without. With startle, the minimum altitude before a missed approach maneuver was initiated was approximately 50 feet below that without startle, indicating startle delayed a decision by approximately 5 seconds. There was observed a moderate correlation of startle effect with age with older pilots showing a greater effect but no association with experience (rank). No simulator metrics other than altitude at decision to go around were reported. This study may be particularly relevant to LEP evaluation in

that a startle stimulus could be an unexpected laser exposure at a critical point in landing or takeoff. However, since there is evidence that startle effects diminish rapidly with repeated exposure, careful consideration would need to be given to the design of an experimental study.

Beer and Freeman (2005) investigated the effect of short bursts of laser light in the central visual field on the ability of pilots to maintain an ideal flight path during night landings using a head-up display in a synthetic cockpit. Their experiment compared brightness matched red versus green and continuous versus strobing laser exposures (8 Hz, 50% duty cycle). Mean flight error and standard deviations of heading were dependent variables used to assess the impact on performance. For flight error, there was a significant effect of laser but post-hoc paired comparisons showed that the only significant laser effect was between no laser and continuous green. All laser exposures impeded the ability to maintain heading and occurred at approximately to the same extent since no differences were found between the laser conditions. Strobing the beam did not t significantly impact flight error performance over baseline. However, both strobe conditions (green and red) disrupted the ability to maintain heading as did both of the continuous laser conditions.

4.3.7 Simulator fidelity evaluation.

In a study comparing two flight simulator software packages, , X-Plane 9 (Laminar Research) and Microsoft Flight Simulator X, Babka (2011) extracted flight performance data for three maneuvers: stalls, steady turns, and flight path stability and compared these results to data obtained from a real Cessna 172SP flying the same maneuvers. The primary comparison shown in the paper were altitude profiles for the three types of maneuvers. The paper contains a good description of how the simulators were set-up to mimic the actual aircraft as well as the test conditions used during the different maneuvers. The conclusion was that simulator software packages produced accurate representations of performance within the actual aircraft and that neither was better than the other.

4.4 Visual Function Studies

Even though flight performance is critically dependent on vision, few studies have directly investigated the linkage between distinct aspects of visual function and performance on different flight tasks. This lack of research was highlighted by Kumagai et al. (2005) in a report that documents an investigation of the basis for the Canadian Forces aircrew entrance vision standard. Numerous visual functions were identified and referenced to performance on flight related tasks. However, because studies on flight performance and a specific visual function were sparse, the authors relied on results from other transportation areas such as driving and shipping to establish links between specific aspects of visual function and performance of tasks similar to those performed in the flight environment. The visual functions identified (and flight tasks they were linked to) included: far acuity (target identification), near acuity (reading), contrast sensitivity (target detection), visual fields (daytime low level flight, peripheral target detection) and useful field of view (taxiing, target detection), glare sensitivity and recovery (takeoff and landing; reading and viewing instruments), color vision (interpreting color-coding, target detection and identification), night vision (night landing, locating runway at night), depth perception (formation flight, landing with crowding), motion perception (hovering over moving ship, detecting air traffic, detecting flashing warning lights). While these linkages between visual function and specific tasks make intuitive sense, few have been experimentally verified in the flight performance domain. In addition, there are other visual functions not mentioned in the list.

Kumagai et al. (2005) also proposed two practical/operationally relevant tests based on near and far acuity since many aspects of flight depended on one or the other. Although they don't provide data to confirm the validity of these tests, the tests themselves provide additional examples of the type that might be used to link function with performance. The proposed near acuity test required locating pre-cued information of varied sizes and contrast on approach plates under four levels of positive sphere blur with accuracy and response time data recorded. The test description did not make it clear how the approach plates would be presented or how accuracy would be determined. The far acuity task was a detection and identification task performed in a simulated approach to a landing or SAR area. Targets were five objects of realistic size varied over three levels of contrast that would appear on digitally photographed backgrounds (e.g. runways, grass, and water) at day and dawn/dusk conditions. The size of the scene is varied to simulate an approach and the subject's task is to first indicate if any of the five pre-cued targets is present in the landing area (target present) or not (target absent), and then as the approach continues and scene size increases, to identify the target. Target size at detection and identification is recorded. The task was to be conducted with best correction and the same four levels of positive sphere blur used in the near acuity task.

Kruk, Regan, Beverley, and Longridge (1983) investigated the association between sensory visual tests and A-10 simulator flight performance in student, instructor and fighter pilots. The vision tests included supra-threshold velocity discrimination of a radially expanding flow pattern, manual tracking of motion in depth and in the frontal plane, motion thresholds and contrast thresholds motion perception and spatial contrast thresholds for grating patterns. Simulator tasks were low level flight, formation flight, bombing, and restricted visibility landing. Figure 16 illustrates one sub-task in the formation flying task; the ideal spacing in the fingertip formation. The dependent variable for this task was total time out of a two-minute formation task where position was maintained within ± 3.6 meters in x, y, or z directions. The other tasks had different dependent variables.

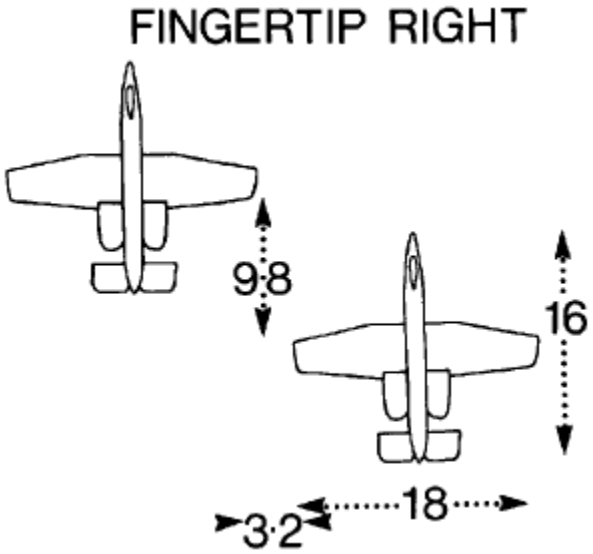


Figure 16. Ideal separation distances (meters) for fingertip formation as well as dimensions of the aircraft. Z-axis not shown (from Kruk et al., 1983)

Several strong correlations between performances involving laboratory vision tests with several aspects of performance in the simulator were found. Tests that were best predictors were the supra-threshold tests of motion sensitivity. In particularly, velocity discrimination of an expanding flow pattern correlated with all aspects of pilot performance. In addition, the manual tracking tasks of motion in depth and in the frontal plane correlated with landing and formation flight. In contrast, threshold tests like motion and grating contrast sensitivity tended not to be predictive of performance, but there was limited variance in test scores for the samples. Unfortunately, the study lacked a matrix showing how the tests themselves correlated with one another and the statistical analysis did not use multiple regression. However, this is an important paper to consider if laboratory vision tests are to be performed in addition to simulator testing. In particular, an optic flow and tracking test in the presence of a glare source could be valuable.

In contrast to Kruk et al. (1983), Ginsburg, Evans, Sekule, and Harp (1982) and Ginsburg, Easterly, and Evans (1983) found that threshold spatial contrast sensitivity at several frequencies was a better predictor performance on the task of target detection . Ginsburg et al., conducted two studies with pilots. The first was conducted in a simulator and involved detection of a ground target (MIG aircraft) during a landing approach to an airfield. (Ginsburg et al., 1982). The second was conducted in a field situation and , involved detection of an approaching aircraft (Ginsburg et al., 1983).

In the simulator study, one pilot flew the mission while a second pilot was in a linked simulator simply viewing the scene. Both responded with button presses on the stick when they detected the target. Only the data for the passive pilot were reported. Landings were completed for three flight conditions of simulated visibility including daytime, nighttime and fog, but only the nighttime landing data were reported. The target aircraft was presented at 37% contrast at the near end of the runway. After completing landings, pilots were shown four photographs of the MIG taken at different distances and asked to choose the one that most closely resembled the appearance of the target when it was detected. Acuity and contrast sensitivity were measured under photopic and scotopic conditions. Spatial frequencies tested with stationary gratings were 1, 2, 4, 8, 16 and 24 cycles per degree (cpd). Frequencies for drifting gratings (5°/sec) were 1, 4, and 8 cpd. The results indicated a relationship between slant detection range and the appearance of the target in the photographs. There were no significant correlations between detection range and photopic or scotopic acuity. In contrast, several significant correlations between contrast sensitivity to low and mid-spatial frequencies and slant detection range were reported such that higher sensitivities were associated with greater detection ranges. The highest correlation was $r = 0.83$ for the peak of the scotopic function.

In the field study, pilots located at the end of a runway were asked to signal when they first detected an approaching aircraft (T-39). Eighty-four pilots were tested over a ten week period. On one day of each week approximately 8 subjects were tested with testing conducted in early morning or late afternoon. Photopic acuity and contrast sensitivity for stationary gratings were measured using the same methods of the previous study. Since meteorological conditions varied each week the data were analyzed by week. The result indicated that mid to high spatial frequencies (8, 16 and 24 cpd) most frequently correlated with slant detection range (higher sensitivity ~ greater detection distances). Visual acuity rarely correlated with detection range and sometimes negatively. In general, the results supported their previous study, however, in the simulator detection tasks, low-

to mid-spatial frequencies were the principal factor while in the outdoor study it was mid- to high spatial frequencies. The differences were not discussed but the cause is most likely related to differences in the ambient lighting conditions which were nighttime for the simulator test and daytime for the outdoor field test. One of the interesting aspects of the study is the discussion of the problems that occur in field testing conducted over extended periods of time. Not only did meteorological conditions vary greatly but so did the color of the T-39 aircraft and how many approaches could be completed during a test session. These factors were out of the experimenters control and resulted in their attempting to control variability by analyzing a given week's data rather than averaging over weeks.

A study that was reviewed previously (Luder & Barber, 1984), is mentioned again in this section since the investigators used a visual search and identification methodology in a dual task structure to evaluate color coding on displays. The goal was to assess the value of color coding in situations where it provided redundant information about the status of a fuel management display. The results were discussed in the context of serial and parallel search. The subject's task was to respond "true" or "false" to statements about the status of the fuel valves. For the monochrome group, the valves states, open, closed or emergency were coded by shape only while for the group provided with color as a differentiator, they were coded by both shape and color. In the search condition the statements were general, e.g., "there are three valves open". In the identification task, statements were specific, e.g., "valves two and six are closed". There were two display sizes with five or nine valves and statements that involved status of one to four valves at a time. The results were discussed in terms of parallel and serial search processing. In the search condition, parallel processing of color resulted in a large response time advantage for the color group, whereas the monochrome group had to search serially because shape was not a salient cue, meaning each valve had to be inspected to determine its status. Color did not help in the identification task because only pre-cued valves at specific locations needed to be inspected. Another finding that suggests parallel processing of color is the lack of an effect of display size on response time of the color group in the search condition while there was an effect for the monochrome group. Finally, the study found that the color group performed better on the compensatory tracking (flying) task, presumably because the presence of color-coding reduced the cognitive load for the fuel status task.

Zárate (2012) used a flicker paradigm to assess differences between experienced and novice pilots, as well as instrument location, in detection of change. In that paradigm, a blank field was presented briefly followed by an unaltered image of a common set of six cockpit instruments then presentation of another brief blank field followed by a picture in which one feature of a specific display was altered. The sequence continued for one minute or until the change was detected. The hypothesis that experienced pilots would detect change more quickly was not supported by the response time data. This result was stated by the investigators as attributed in part to low statistical power, but also to the lack of inclusion of a non-pilot control group. The other hypotheses including that change would be detected faster on the ADI because it is the most important instrument and centrally located in the most common T-scan pattern was supported by the data. Inaccuracy and trial data over time were also analyzed and no significant effects were found. Experience was one of the dependent variables, however, measures of visual function such as acuity or contrast sensitivity or visual search performance could also have been used.

The last study in this category to be reviewed is that of Schmeisser, Maier, Freeman, and Brockmeier (2005). Although the study relates to LEP evaluation and laser exposures, and effects on vision rather than relating visual function and pilot flight performance, the methods used are useful for design consideration of studies assessing LEP effects on flight performance. In their study, pilots' visual acuity was captured using a HUD as the display, with and without LEP and with and without laser glare present. The general purpose was to investigate the cost-benefit matrix associated with laser eye protection worn in the cockpit and more specifically to determine if there were differences in cost-benefit between different LEP technologies and as a function of the level of protection. Four measures were derived to evaluate LEP as they related to visual effects (Figure 17).

The first measure was cost, defined as the decline in vision with LEP compared to no LEP when no laser was present. The second, protection, defined as the gain in vision with LEP in the presence of laser glare. The third was performance, the difference in vision between a no LEP no glare condition and LEP with glare condition. The fourth was efficacy, which was the difference in acuity with LEP between the glare and no glare conditions. One finding that all LEP have some cost in that they reduce acuity was expected. It was also expected that acuity would improve in the presence of laser glare as optical density (OD) of the LEP increased (protection) and that was verified. Also, the higher-level OD's had a negligible cost compared to lower levels that resulted in higher efficacy for the higher OD LEP. The reflective technologies had better efficacy than the dyes, but were rated less desirable by the subjects, particularly with glare present. This latter finding is consistent with other studies that have found differences in effects on vision between dye and reflective LEP with glare present. The findings have been related to differences in light scattering properties of the two types of technology (Dykes et al., 2004; Kuyk, Smith, et al., 2013).

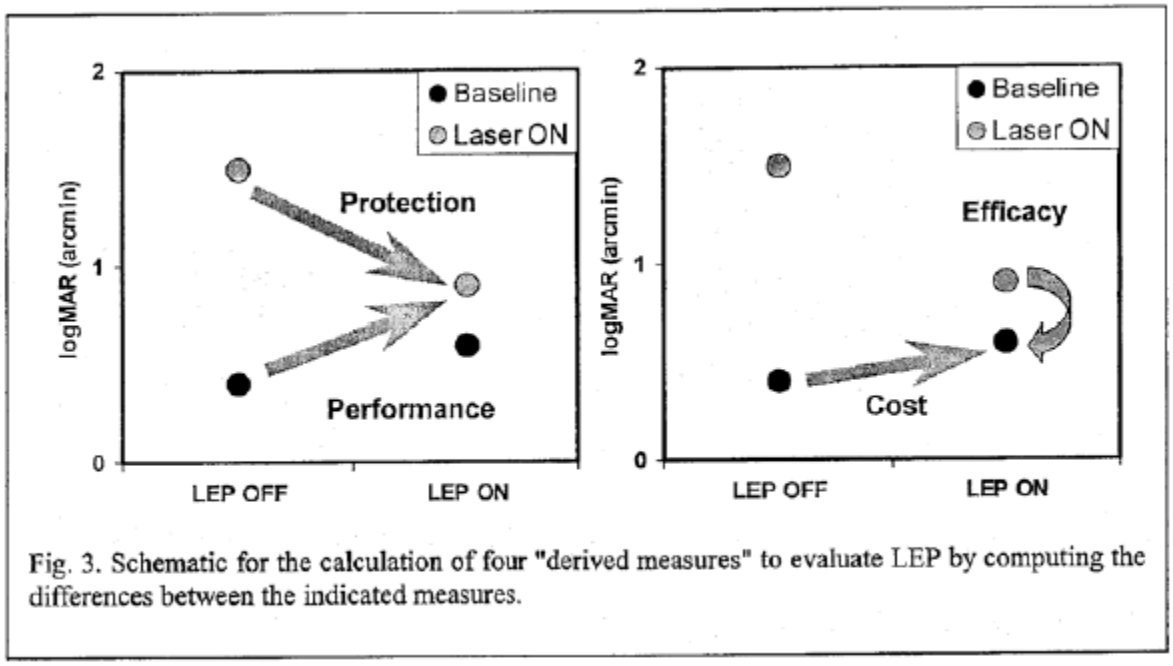


Fig. 3. Schematic for the calculation of four "derived measures" to evaluate LEP by computing the differences between the indicated measures.

Figure 17. Figure 3 from Schmeisser et al. (2005), showing the four measures used to evaluate LEP

4.5 Miscellaneous

Several studies did not fall into one of the three categories used to group papers for this review. Nevertheless, they provide valuable and relevant information on assessment methods that might be incorporated into evaluations of LEP.

Leland, Rogers, Boquet, and Glaser (2009) conducted a study similar to the startle study of Martin et al. (2012) in that during testing in real aircraft flight, pilots were challenged by placing them into upsetting conditions, which is analogous to being presented with a surprising/startling situation. Their study concept involved training in simulators to reduce the startle aspect. There were two training groups in the study and one no training control group. The two training groups received ten hours of upset recovery training in two distinct types of simulators: a high fidelity full-motion system versus a low fidelity desktop simulator running MSFS. All subjects then performed upset recovery tasks in an actual aircraft. Data were recorded by analyzing video from the control panel and from a flight recorder system. The three main conclusions were that the trained groups out-performed the untrained groups there was no significant difference due to the type of simulator used for training, and both trained groups were judged to fall well short of upset recovery skills of aerobatic pilots. The last suggests that while simulator training is useful, it is not be a complete substitute for training and flight experience in a real aircraft.

Wei, Zhuang, Wanyan, and Wang (2013) used the situation awareness global assessment technique (SAGAT) developed by Endsley, Selcon, Hardiman, and Croft (1998) and Strater and Endsley (2000) to assess situation awareness for three commercial airline cockpit display interface (CDI) designs. The purpose was to demonstrate that CDI could be evaluated for SA in the experimental flight simulation environment during the design phase of display development. The SAGAT is a computerized memory probe measure, and Wei et al. (2013) used the “freeze” technique previously described by Diez et al. (2001) to implement it. In a randomized-block design thirty subjects with extensive simulator experience monitored the CDI during a flight scenario that involved take-off, cruise, and landing. At random times during the flight the CDI froze and was replaced by an interface that asked specific questions from the SAGAT about flight status, which subjects had to recall from memory. Heart rate was also recorded as it has been shown to be associated with cognitive workload. Differences were found between the three displays with one providing better SA but also requiring higher workload to do so. The interesting aspect of this study is the use of subjects monitoring displays but not flying the aircraft and the freeze technique to assess situation awareness. Both methods could be applied to interpretation of changing and dynamic information presented on color displays with and without LEP. Heart rate might also be monitored to determine if the alterations in display appearance by LEP effects cognitive load.

A presentation by Ford and Munro (2011) outlined the basics of the MAPP (Model for Assessing Pilot Performance) developed by Mavin and Dall'Alba (2010). The MAPP model consists of interrelated essential skills of situation awareness, flying performance, decision making, and factors that feed into those, including knowledge, management, and communication. The model presented was purported to be simple to understand in a way that will allow pilots to self-assess but can also be used by examiners. It was s based on a hierarchy of skills and allows for integration of technical and non-technical skills. The project outlined was s to develop and implement a training program on use of the model for pilots and examiners.

A different process for assessment of pilot attention is described by Cannavò, Conti, and Di Nuovo (2016). Rather than measuring eye movements or pilot recall using the "freeze" technique, it consists of a battery of seven computerized tests to measure selective and divided attention. Tests in the battery include simple reaction time to centrally and peripherally presented visual stimuli, multiple search plus memory, color-word interference based on the Stroop test, a ground interference test involving discrimination of a target in an active background, divided attention with auditory and visual "targets", digit-span memory, and a "global vision" task involving detection of moving stimuli among moving distractors. To validate the battery, it was administered to experienced pilots and untrained controls, with experienced pilots presumably having better attention skills. The pilot group showed better performance on all tests, and discriminant analysis indicated the tests could discriminate between the two groups with a high level of accuracy. Additional step-wise regression analysis identified the core subset of variables that were included in the final model. Most of the final model variables were measures of errors rather than time factors. Two variables, multiple search plus memory and divided attention, did not contribute to the model. One criticism is that the tests are no more than modifications of existing tests and although they are based on the literature, full references, other than to the Stroop test, were not provided. Nonetheless, similar tests may be useful in LEP effects evaluations, particularly those involving discrimination of targets against varying backgrounds using central and peripheral vision.

4.6 Laser Eye Protection and Visual Function

As noted previously, LEP works by preventing laser light from reaching the eye or reducing its intensity to levels that will not cause damage. Although lasers operate at a single wavelength, due to technology limitations, blocking of laser light by LEP is not selective to a single laser wavelength. Rather, a band of the spectrum around each laser wavelength is blocked. When the blocking band(s) occurs in the visible spectrum, the spectrum of light available for vision is altered and this can result in significant effects on visual function. Studies of the effects of LEP on vision have concentrated on color and spatial vision. These variables are investigated due to their importance for flight performance and because they are likely to be affected by changes in the spectral content and the total amount of light available for vision. In contrast, the effects of LEP on, for example, motion detection, velocity discrimination, optic flow, visual tracking, search, and depth perception have not been studied in detail.

Although there is little quantitative data on LEP effects on visual function other than color and spatial vision, comments and responses to questionnaires made by pilots during ground and flight testing of different LEP suggest that other visual functions are also adversely effected by wearing LEP. In a recent test of a prototype LEP that had similar tint characteristics to shooters glasses, several rotary wing pilots and crew stated that the contrast enhancing effect resulted in difficulty judging altitude during low level flight over a desert environment (Novar et al., 2015). The contrast enhancement made objects appear closer than they really were and, if only visual cues were used, resulted in positioning the aircraft higher above ground than desired during hover/sling rescue operations. For the same LEP, the yellow tinting led to difficulty judging the distance of far objects due to contrast reduction and blending of objects against backgrounds. Both observed effects suggest some LEP may impair depth perception in situations where contrast provides an important depth cue. Laboratory tests of stereopsis, however, have not found significant effects of LEP, likely because these tests involve high-contrast achromatic stimuli that are largely unaffected by wearing LEP (Schmeisser et al., 1999).

Another common comment from new LEP testing is that because the LEP reduced the overall light level inside the cockpit they increased the problem of transitioning from a bright outside scene to a dim inside scene and vice versa. Both transitions take time for the pilot's visual system to adapt to the new light level, regardless if LEP are worn or not. However, a possible reason why LEP seem to exacerbate the problem could be related to the LEP increasing the effective range over which light and dark adaptation needs to occur in a daytime setting. Take for example LEP that has a photopic luminous transmittance (PLT) of 40%. In other words it reduces the amount of light useful for vision y 60%. For a bright outside scene with luminance levels of around 5,000 $cd \cdot m^{-2}$ the LEP will reduce that to 2,000 $cd \cdot m^{-2}$, which is still well above the level where acuity and contrast sensitivity reach maximum levels (~300 $cd \cdot m^{-2}$). In contrast for a cockpit shadowed interior at around 200 $cd \cdot m^{-2}$, the LEP will reduce the level to 80 $cd \cdot m^{-2}$, which is in the range where acuity and contrast sensitivity are lower than maximum levels and require some dark adaptation time regardless of the direction of the transition.. Thus without LEP, the transition from outside to inside occurs in a light range where the visual system is functioning optimally or close to it. Whereas with LEP the transition is from an optimal function to a reduced level, or vice versa, that pilots report as problematic.

Other comments about the overall reduction in light levels with LEP have been that they are too dark for night operations or too light for high brightness environments such as in desert environments or over water on clear days. Even though photopic luminous transmittance (PLT) levels for night specific LEP have been increased over the years by virtue of advancements in dye and thin film coating technologies, in a recent flight test of an LEP with PLT just above 50%, questions arose about their safety during night re-fueling under blacked out conditions (Putnam et al., 2017). It has been suggested that increasing PLT to levels above 59% might resolve this type of problem for night specific LEP (LaFrance, Kent, Foutch, Miller, & Kuyk, 2009). However, that value may need to be even higher, since in that study, night flights were conducted under optimal conditions; cloudless sky and full moon. In contrast, in a laboratory study, Martinsen, Havig, Dykes, Kuyk, and McLin (2007) found little drop off in performance on an acuity task until PLT declined to approximately 50%. The discrepancy between studies leaves the PLT for night LEP question unanswered and further defining the effects would be desirable through investigation in simulators that have the capability to simulate night and low light flight conditions.

The opposite problem of day specific LEP being too bright for some daytime environments can and has been addressed by reducing PLT with neutral dyes and providing aircrew with several levels of PLT to select from (Belleau, Mariano, Novar, Chung, and Kuyk, 2015). Unfortunately, current flight simulator platforms do not have the capability to generate light levels anywhere close to bright daylight, so the adequacy of a PLT level for use in a high brightness environments cannot be addressed in them. .

A third common comment is that color shifts or contrast reduction make information on displays more difficult to interpret and result in a slowing of processing speed. For example, with a yellow tinted LEP, white colors on displays are shifted toward yellow. Usually the hue of the shifted white does not appear exactly the same as a real yellow so the two can be discriminated. However, because the color difference is small, the discrimination takes more time. Also, the two stimuli are often not side by side for comparison or present at the same time so retrieval from memory of their attributes must be relied on to make the shifted/real judgement. Results from a color identification test with LEP support the slowing of processing as several studies have found response times

increase significantly with LEP and that the increase cannot be explained on the basis of light reduction (Kuyk, Engler, Brockmeier, Kumru, & Mariano, 2013). Similarly, contrast reduction of a stimulus can slow processing speed since response times generally increase as stimulus contrast decreases. One note about contrast is that there are two types: luminance and color. For the most part, contrast based on luminance differences is perceived as a difference in relative brightness between a target and the background. Color contrast occurs between two different colors of the same luminance and varies depending on similarity of hue, but also varies depending on apparent brightness. Luminance contrast has several quantitative and accepted definitions but color contrast does not.

5 DISCUSSION

5.1 Summary and Suggestions for RHDO Simulator Studies

For commercial and general aviation, all flying involves multiple elements including maintaining heading, attitude, altitude, and airspeed during different phases of flight from takeoff, climb, cruise, descent, approach, landing, and taxiing, but also includes altitude changes, course changes, and aerial maneuvering to avoid air traffic or obstacles. Other tasks are detection, identification, and interpretation of information provided on instruments, displays, maps, and approach plates and detection and response to caution and warning lights, detection of air traffic, obstacles, and landmarks.

In the literature reviewed, flying performance was assessed on a variety of different flight tasks with approach and landing being the most frequently used, but studies also included takeoff and cruise phases, navigation between waypoints, turns and other maneuvers and avoidance of air traffic and ground obstacles. Flying performance was assessed by extracting data from simulators and either using it in raw form (e.g. altitude at runway threshold), or processing it to determine deviations from ideal flight parameters (e.g., heading, altitude, airspeed, and glideslope) or to indicate average or median performance. Some studies combined multiple measures to provide indicators of overall flight performance. Also, frequency and amplitude of inputs to flight controls have been used as indicators of performance. The simulators used in the studies reviewed within this document ranged from simple desktop systems with a single display and a stick control to full motion, high fidelity systems. Subjects tested ranged from those with no flight experience to highly experienced commercial airline and military pilots.

In addition to flight metrics derived from simulator parameters, interaction with displays, instruments and maps has been assessed using eye movement metrics and the concept of AOI's to assess attention distribution and SA in response to changes in display properties, in-flight emergencies, phases of flight, age and expertise of pilots. Other measures of SA included self-report of information content of displays and flight status using a "freeze" technique, and physiological metrics including changes in pupil size, heart rate, and blood oxygenation as indicators of cognitive load or workload. Detection of obstacles, hazards, and warning lights has generally involved the metrics of response times and errors (omissions), although in some cases results have been expressed as distances in slant range.

For military aviation, the range of flight tasks performed is larger and they may be performed under time of stress or requiring a rapid activity cadence in critical circumstances. For example, target acquisition, tracking, aiming, high speed maneuvers and weapons release or formation flying are not tasks civil aviators are likely to perform. Despite this, the metrics for assessing flight performance are like those used in civil aviation studies. They consist of raw measures of parameters like altitude, speed or bank angle, combinations of parameters, averages, deviations from an ideal flight path, and input to controls. Similarly, interaction with displays has been assessed with eye movements or by self-report, and detection of targets or hazards with both manual and verbal responses. The use of NVGs in flying is unique to military operations, but discussions relating to interactions between NVGs, LEP and flight simulators is beyond the scope of this report.

Considering the review of flight simulators and other comparative studies compiled here, there are a number of take-away approaches for testing LEP performance in flight simulators as a step in the process of determining flight functional use and pilot acceptance. Since the primary changes in visual stimuli caused by LEP involve color, contrast, and reduced light levels, those should be the focus of LEP evaluation. Color changes or loss due to filtering of visible light by LEP can be expected to have effects on extraction of information from color coded displays or other lighting, e.g. PAPI/VASI (Visual Approach Slope Indicator) and warning lights, but also on detection of targets and hazards. Changes in contrast can have similar effects and most likely are the result of the color intensity changes that occur with LEP. For example, the contrast of a bright color on a dim background can be reduced if the LEP absorbs a significant amount of the light in the spectral band of the colored stimulus and reduces its intensity while having a much smaller effect on the background. However, contrast changes with LEP can also enhance the stimulus, making it easier to detect. A real-life example is the increased contrast of a grey target against a blue sky background when they are viewed through LEP that selectively block short wavelengths. In this case the LEP significantly darkens the blue background but not the grey target which effectively increased the contrast between the two. The overall reduction of light level caused by LEP is likely to result in a general decrease in performance, particularly when LEP are worn in conditions where light levels are low.

For the RHDO simulator, the possible approaches for testing LEP described in the following paragraphs are in the context of using non-pilots as research subjects. The potential approaches include some that do not require subjects to fly the simulator but rather act in the role of a spotter or co-pilot, whereas others involve having novice subjects execute simple flight tasks. The latter will require some training, however, the MSFS software that controls the RHDO simulator has flight training lessons as part of the package. While there are limitations imposed by using non-pilots, useful and important information with respect to LEP effects on visually guided/dependent tasks can still be obtained. Many of the tasks performed by pilots are performed in everyday life by everyone else (e.g. target detection and identification) or can be structured to require minimal training to reach a level of proficiency sufficient to measure LEP effects (e.g., executing an instrument cross-check).

A research study proposed several years ago for the 711HPW Research Studies, and Analysis Council (RSAAC) program still has relevance for LEP evaluation in a simulator environment. The study concept involved training subjects up to a criterion level on a repetitive instrument/display monitoring task and then having them perform the same task with and without LEP. If LEP has an adverse effect on extraction of information from displays and instruments the expectation is that the time to complete the monitoring task would increase and more errors would be made. The monitoring task would involve reporting of information on the displays/instruments during a flight scenario and possibly interacting with some instruments in response to certain changes. In the original proposal, the subject would not be flying the simulator but would be performing a sequential monitoring task from the pilot's seat because it provides optimal viewing of the instruments. . For consistency, the scenarios during which subjects perform the monitoring tasks would be pre-recorded flights that are played back through the simulator in a random or counterbalanced sequence. This random flight presentation would ensure all subjects are challenged with the same information but with changing flight conditions.

Since the LEP simulator does not have a glass style cockpit with many color-coded displays it may be necessary to extend the monitoring task beyond a display/instrument task. This additional challenge in the paradigm could include the random introduction of caution and warning lights (with no auditory component) or pop-up hazards such as radio towers or other air traffic that subjects must respond to either verbally or with button presses. It might also include monitoring OTW components to report status of PAPI/VASI lights or tower warning lights during a landing approach, detection of hazards, such as other aircraft in the air or on the runway, and the detection and identification of landmarks or other ground targets. Whether to combine OTW tasks like monitoring PAPI/VASI status with in-cockpit monitoring, or to treat them as a separate task may require pilot studies to determine how non-pilots respond to them.

Another additional challenge would be to put a secondary color display in the cockpit that subjects must monitor and use to perform a color related task, such as presence of friend/foe in a designated sector of the airspace (Gaska, Wright, Winterbottom, & Hadley, 2016) or a monitoring task like the fuel management task used by Luder and Barbur (1984). The secondary display might be a small notebook or laptop, possibly with a touch screen display to facilitate response input. Integrating the notebook with the simulator should be achievable so that the two systems communicate with each other and coordinate the timing of the tasks performed with each system.

Integrating eye tracking capabilities into the RHDO simulator should be considered as a near term goal. Eye movement metrics could be used to assess LEP effects in the monitoring task described above, particularly as they relate to distribution/allocation of attention and information processing. These capabilities would be particularly useful for determining if LEP alter scan patterns due to changes in display colors (shifts) and symbology contrast, but also to provide additional information about performance on tasks like target detection and tracking. Eye movement metrics could augment response time and accuracy data obtained on the cockpit monitoring task and allow more accurate determination of where in the cockpit problems with LEP occur. A variety of metrics can be extracted from the eye movement data and a good description of many of these can be found in Moacdieh et al. (2013). However, the analysis of dwell/gaze times across a set of AOI's and number of fixations in an AOI that includes the cumulative fixation duration and mean fixation duration have been the most commonly used.

Another task that could be performed without having subjects fly the simulator is target detection and identification. In an airborne target detection task the simulator would be flying a pre-set course along a straight and level flight path at fixed speed. Aircraft targets would appear to the left or right of center, at different eccentricities and altitudes, but along a heading at fixed speed that would have them pass in front of the flight path of the simulator at some pre-defined distance. The airborne targets could be different types of aircraft, fixed or rotary wing (or maybe colors) and presented against different backgrounds (blue sky, clouds, terrain). The subject's task would be to respond when they first detect the aircraft by identifying its location relative to the simulator flight path, left or right or possibly in one of several pre-defined zones. The subjects would respond again when they can identify the type of aircraft (fixed or rotary wing) or its color. When the aircraft first appears they will be of a size that is below detection threshold as determined by pilot experiments. As the target aircraft and simulator converge, the target will increase in size as it would in a real world situation.

For the ground target detection task the simulator will again fly a straight and level flight path at fixed speed, but at a lower altitude, or possibly on a gradual descent with the nose pointed slightly downward to increase the view of the ground to the subject. Ground targets could be different types of vehicles, an aircraft either on the runway or the taxi apron (e.g., Ginsburg et al., 1982), or some other target such as smoke or signal lights of different colors. They could be presented on the same background or different backgrounds (woodland, fields, desert, possibly even water) and at different locations relative to the subject's view in the simulator. The subject's task is the same as for the airborne targets, detection and then identification. Simple flying tasks are also a possibility and the feasibility of having them performed by non-pilots should be explored. The RHDO flight simulator uses the Microsoft Flight Simulator software that includes flight lessons designed to help learn how to control the aircraft and perform tasks like take-off and landing. It may be possible to use these lessons to not only train novice subjects to simulated flight, but also for key data gathering. For example, MSFS, has a flight lesson where virtual hoops appear in the airspace though which the pilot flies the aircraft. The software provides feedback on performance relative to an ideal flight path and those parameters could be used as dependent variables. This flight lesson was used by Rossi et al., (2006) to explore if the use of out-the-window cues provided any advantage for landing and level turn training.

Simple flight tasks might be combined with the in-cockpit monitoring task described previously or set-up to require some monitoring of instrument to maintain flight status while responding to unanticipated events. Several possibilities are inspired by the reviewed literature. One task might involve a simple tracking task, such as keeping a leading aircraft flying at constant speed but variable course within a reticle (sight window) placed inside the simulator and then comparing flight paths for the target and test aircraft (Kasarskis et al., 2001). Alternatively, the tracking task might simply involve following another aircraft from a close distance, like the formation flight following task described by Kruk et al. (1983). This type of task would be used primarily to increase the subject workload for potentially enhancing LEP effects on the monitoring task since wearing LEP is not likely to influence tracking or following performance per se unless the tracked target is very small and its contrast significantly reduced by the LEP.

Another simple flying task might be to have subjects fly a straight and level course at constant speed complicated by varying wind speed and direction so that they must actively work to maintain heading, altitude, and airspeed. This paradigm would require the test subjects to monitor certain instruments and extract information from them. Here again, wearing LEP is not likely to influence flight performance unless the PFD contains color-coded information whose visibility is adversely affected by LEP. However, to gauge LEP effects, caution or warning lights could be activated or color-coded or non-color-coded stimuli presented on the PFD or in the outside environment at different eccentricities and have the subjects detect and respond to them in real time (Russo et al., 2004). A non-color-coded stimulus might be the introduction of another aircraft, either in the air or on the ground that subjects must detect, or detect and then identify aircraft type, flight direction or location relative to a runway being approached (on it or not). Time after stimulus onset and detection/identification range could be the outcome measures.

5.2 Challenges

Any flying task, even a simple one using limited flight controls will require training non-pilot subjects to fly the simulator. However, this need not be exhaustive training if the task is simple (Chuang et al., 2013; Khan et al., 2006). In addition, subjects could be pre-screened for flying

aptitude by having them fly a straight and level course at fixed speed and altitude with variables like mild turbulence or cross-wind added to force them too continuously and actively control the aircraft. Subjects with the smallest deviations from an ideal flight path would be used as subjects (Khan et al., 2006).

Potential problems with monitoring subject performance and presentation of flying tasks mentioned is that the RHDO simulator may not be capable of being programmed to introduce specific color stimuli in the outside environment or stimuli that behave as real objects. For example, objects that increase in size as they are approached is a prerequisite to an accurate detection range task. Similarly, introducing a color monitor that is not a normal part of the simulator and using it to display information subjects must respond to would require linking it to the simulator so that the timing of a stimulus presentation can be linked to the timing of flight parameters or control settings.

There are also issues with using the color capabilities of the OTW scene to simulate real-life stimuli like PAPI/VASI lights that need to be considered. For example, the spectra of older PAPI lights is broad band since it is generated by filtering an incandescent light source with a color filter. In contrast, a PAPI stimulus generated with a projection system like that of most simulators will be a mixture of the narrow band three color (red, green and blue; RGB) spectra of the projector. Using a color monitor has the same problem. While the simulated PAPI lights may match the real lights in appearance (metamer) they will have a different spectral energy distribution and as a result will not be filtered by LEP in the same way as an actual PAPI light. This difference could result in color appearance changes caused by the LEP between the actual and simulated lights that may result in differences in subject performance on a PAPI light task. The same, and potentially worse problem, could occur in trying to simulate newer PAPI lights that contain LED's. Red LEDs put out a spectrally narrow band of light that would be difficult to reproduce if the emission spectra of the red (R) monitor component is not reasonably similar. White LED's may be no less problematic if the white is created by mixing light from several different LEDs. Even though their output is a mixture and spectrally broader, a monitor reproducing them would need to have its red, green, and blue (RGB) output spectrally close to the output of the individual LED. Phosphor based white LED's may be less problematic since they have a reasonably broad spectrum that may be easier to approximate with a mixture of display/projector RGB components.

Using a commercial display inserted into the simulator to replicate real cockpit displays has the drawback that the spectra of the R, G, and B components that are mixed to generate a color palette are not likely to be the same as those on military displays. Thus, stimuli generated on the commercial display, while being metameric matches to stimuli on military displays are not spectrophotometric matches and like the PAPI light example, their color appearance through LEP may not be the same. However, there are also drawbacks associated with using displays taken from cockpits. For one, there are diverse types of cockpit displays made by different manufacturers so that selecting one for test purposes and using it to duplicate symbology and colors for another display will likely yield similar problems with metameric matches as those for commercial displays. Secondly, military displays removed from the cockpit are also removed from the systems used to drive them and generate the symbology and colors. What is often required is development of a system for driving the display and controlling its output. This can be a time consuming and complicated process making it unfeasible, due to cost and/or time, to assemble a variety of cockpit displays for laboratory testing.

Despite drawbacks, the use of commercial displays for LEP effects evaluation is still likely to yield generalizable results (Gaska et al., 2016). Also, it may be possible by careful comparison of the spectral output of military displays, if that information can be obtained, to find a commercial counterpart that is either close in properties to a display or interest or approximates the properties of several cockpit displays.

There are challenges facing the use of eye tracking systems in simulators. One will be to determine if eye tracking systems that rely on near infrared (NIR) imagery, and most do, will work with LEP or even work well enough with clear lenses imposed in the image path. The problem is that many military LEP block NIR light. Most eye tracker work by illuminating the eye with NIR light that is then imaged by NIR sensitive cameras and used to locate the pupil and certain reflections in the eye. The NIR illuminators are often located facing the subject with the cameras below and facing the eyes but some distance away from the face. Interposing a NIR blocking LEP between the eye and the illuminator and cameras would block both the incoming and any reflected NIR signal, and therefore, could not be tested with NIR-based trackers. This issue might be circumvented by obtaining LEP samples without the NIR blocking technologies.

Several studies reviewed also reported problems recording from subjects who wore eye glasses. Although the EyeLink II eye tracker owned by RHDO is advertised to work through glasses, assurance of this or if an issue remains would need to be determined empirically. Other factors that could interfere with eye trackers is the ambient lighting condition since one study reporting problems with their tracker in high brightness environments. While this is not likely to be a problem in the RHDO simulator because of its modest light output, it could present a problem for studying eye movements with or without LEP in the presence of broadband or laser glare. Bright light reflected from the face, or from reflective LEP, back to the tracker cameras may overwhelm them, which would be similar to the same problem reported in one study with high ambient light conditions.

Finally, several studies mentioned calibration maintenance as a problem with head-borne trackers due to the large, frequent, and often rapid head movements made by pilots causing the tracker securing straps to shift on the head. This problem is not likely to be solved by using non-head-borne eye trackers since most rely on NIR imaging. Both the NIR and eye tracker shifting problems might be eliminated by using trackers that are in a spectacle configuration (e.g., Tobii and ASL) or trackers that use the electro-oculogram (EOG) signal to establish eye position and movement (Biopac Systems Inc.). The Tobii system, for example, is contained in an eye glass frame possessing NIR illuminators and cameras located very close to the wearer's eyes that is placed behind a protective and removable lens. It may be possible to modify LEP to serve in place of the protective lens. The spectacle eye tracker configuration removes the NIR blocking problem and the problem of having a lens between the tracker cameras and eyes. Furthermore, calibration loss due to shifting on the head may be obviated if the spectacle trackers are constructed with significantly lighter weight materials and do not have arms containing cameras extending from them, which impose angular forces on the straps when the head is quickly moved. The spectacle frames sit close to the face and movement of them should be minimized if well fit to the nose and snug on the heads of subjects. A suggestion here is to determine whether the sensitivity of spectacle systems like the Tobii and the data processing software available for them are adequate for research needs. Systems based on EOG should also work with any LEP but drawbacks include lower resolution and that subjects must wear electrodes that need to be consistently applied.

Furthermore, there are not many commercially available EOG based systems for eye tracking. Using a standard EOG system may require development of data analysis software, unless it can be obtained from other investigators or is available from sources like PsychToolbox.

An important consideration for any eye tracking system is the ease of defining AOI's in the software that controls them. Having a set of defined AOI's is necessary for determining if wearing LEP alters scan patterns as well as for identifying locations where problems are occurring. Some systems allow AOI's to be pre-defined in the software (SMI), others may necessitate defining them after the data are collected. Without AOI's, the experimenter simply has a set of data that indicates number and duration of fixations and saccades for a specific time. While these data have general usefulness, the data does not allow determination of specific scan patterns and if those change because of an intervention such as LEP. Another consideration for spectacle type trackers is that they generally have lower data sampling rates, in the range of 120 Hz compared to desk based or head-mounted trackers like the EyeLink II or ASL that have sampling rates of 500 Hz or more. The lower sampling rates may reduce the resolution of some parameters such as saccade velocity, saccade duration and fixation duration. Fixation duration is a likely parameter to be used for analysis and it would need to be determined if the small loss of resolution is acceptable. Other factors like number of fixations in an AOI and direction and length of saccades as well as establishing scan patterns should not be impacted.

Testing with qualified general aviation pilots would expand the complexity of flight tasks used for LEP evaluation. These individuals would be able to execute tasks such as take-off, approach, and landing as well as more complicated navigation tasks under full simulator conditions. The seasoned pilots would also have experience with cockpit monitoring procedures and how to utilize information presented on different displays and instruments. No flight training would be required either, just familiarization with the simulator and how it performs its functions, which could be achieved by having practice flights. The disadvantages would be recruitment and scheduling, insuring their experience levels are similar, and likely higher costs to compensate them for their time. Even if pilots are not used as subjects, at least one pilot familiar with the types of aircraft that the RHDO simulator could emulate should be contracted as a consultant to provide expert advice on the design of flight scenarios used for LEP evaluation. If replication of tactical display symbology on a remote monitor is planned, then a military pilot with knowledge and experience should be consulted or one could seek other investigators with developed software that could be contacted about collaborating and/or sharing their programs.

Beyond the RHDO simulator, testing should eventually be conducted with military pilots in military aircraft simulators. The testing would involve comparing performance on established tasks used for pilot evaluation with and without LEP. Even if the simulator cockpit displays suite is not an exact replication of a real cockpit, it should still be possible to gather information about effects of LEP on many tasks such as detection and identification of air and ground targets and detection and identification of low contrast and even color-coded information on displays. Another approach is to test LEP that have already been ground and flight tested in simulators for the same aircraft. Not only could flight performance measures be obtained but also responses to the same checklists and questionnaires could be compared to determine if simulator data are representative of those obtained in the real aircraft. If simulator testing proves to yield useful information, simulators could then be used to test prototype LEP or LEP that are entering the first stages of acquisition in a range of aircraft simulators that represent the aircraft the LEP will most

likely be used. Lastly, if the spectacle eye trackers prove to be effective with LEP in the RHDO simulator, these devises could be used with pilots in military simulators performing combat related tasks. Although eye tracking data exists for some combat related tasks (Yu et al., 2014; Yu et al., 2016), none has been collected while wearing LEP.

5.3 Recommendations for RHDO Simulator Studies of Visual Function and Flying Performance

An extended use of the RHDO simulator system could be for studies that determine the relationship between laboratory tests of vision and simulator flying performance. Visual functions assessed might include spatial contrast sensitivity, temporal contrast sensitivity, velocity and optic flow discrimination, tracking in depth and different assessments of color vison. The simulator can be programmed to present military flight activity elements such as target detection and identification ranges, detection of moving targets, detection of heading changes, direction of flight of other aircraft from wing and tail light colors and configurations, and change detection involving instruments and displays involving color and contrast. These tasks could possibly be performed under simulated settings to emulate day and night conditions, and for daytime, possibly in the presence of a broadband glare source to simulate the sun.

Many of the in-simulator detection and identification tasks could be performed by a passive observer acting as a spotter or it may be possible to train non-pilots to execute simple flight tasks and relate performance on them to visual function. For example, one task used by Kruk et al. (1983) was a formation flight task that required pilots maintain a position within specific limits relative to another aircraft, either in a wing-tip or following task. Others may be maneuvering through the MSFS virtual hoops or performing a simple compensatory tracking task.

The reason for conducting studies that link visual function and flight performance is to develop a comprehensive model of visual function as it relates to flight performance. As Kumagai et al., 2005 pointed out, such a model does not exist and also that few aspects of vision have been directly related to flying performance. The few studies that have been done in this area were never replicated and in some cases used stimuli and methods for assessing visual function that can be improved. For example, (Ginsburg et al., 1982) measured spatial contrast sensitivity using sine wave gratings and a method of limits procedure. The extended gratings could be replaced by Gabor patches for the stimuli and criterion free psychophysical methods such as forced choice. Furthermore, there were limitations imposed by the simulator in one study that required a fixed order of visibility conditions from clear to low that may have resulted in a mixture of visibility and practice effects. It is also not known if the simulators in earlier studies presented the scene in color or black and white (Kruk et al., 1983).

The main limitation of the types of studies proposed is that non-pilots would be used, at least in the initial phases of model development. This approach will result in limiting the types of flight tasks that can be performed. Even if non-pilot subjects were trained to complete tasks such as takeoff and landing, for example, they may not behave in the same way as experienced pilots who would have greater knowledge of the aircraft, its controls, instruments, and behaviors. Even with this limitation, many flying tasks could be performed by non-pilots and their study result performance validity is not likely to be influenced by the lack of flight experience.

Other laboratory-based studies could be conducted using operationally relevant tasks and either actual aircraft displays or commercially available color displays set-up to present current symbology or replicate a set of cockpit instruments. As reference points, Reddix et al. (2014) used a commercial display to present pairs of colored missile symbols on a display from an actual military aircraft. The subject's task was to respond if a target pair (green missile symbol over a red missile symbol) at a specific orientation presented among other pairs of red and green symbols at different orientations was present or not. In a different study, Gaska et al. (2016) simulated 5th generation military symbology on an EIZO color monitor and subjects responded yes or no if an enemy aircraft was located within a set of color-coded boundary lines. While using military symbology insures that the task will be operationally correct, both tasks are essentially serial search tasks and the display content could be simplified and the tasks would still maintain relevance. For example, in the Gaska et al. (2016) task, red and magenta boundary lines defined the search area for the subject who first finds the search area, then scans it to determine if it contains enemy aircraft or not. The aircraft were represented by small aircraft symbols color coded red for enemy, green for friendly and yellow for neutral. This serial search scenario could easily be duplicated using simplified symbols such as circles, squares, or triangles since the task involves search for a specific colored stimulus in a restricted search area. Symbol type was not a search relevant parameter. Similarly, the task in the Reddix et al. (2014) study could have used short oriented pairs of line segments as the stimuli, rather than small missile symbols without impacting the results.

6 REFERENCES

Adams, B. D., Davis, S. A., Brown, A., Filardo, E.-A., Thomson, M. H., & Wood, D. (2013). *Post traumatic stress disorder (PTSD) in emergency responders scoping study: Annotated bibliography*. (DRDC-RDDC-2014-C18). Ottawa, Ontario, Canada: Defence Research and Development Canada - Centre for Security Science.

Ahlstrom, U., & Dworsky, M. (2012). *Effects of weather presentation symbology on general aviation pilot behavior, workload, and visual scanning*. (DOT/FAA/TC-12/55). Atlantic City International Airport, NJ: FAA William Hughes Technical Center.

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics, 36*(7), 623-636.

Appel, B., Schubert, E., & Hutting, G. (2012). *Assessment of manual flying skills in full flight simulators.* Paper presented at the European Airline Training Symposium, Berlin, Germany.

Babka, D. (2011). *Flight testing in a simulation based environment.* (Bachelor of Science), California Polytechnic University.

Beer, J., & Freeman, D. (2005). *An initial examination of visual disruption in a synthetic flight control task, resulting from laser glare presented with different wavelengths and temporal profiles*. (2005-03). Brooks City-Base, TX: Naval Health Research Center Detachment, Directed Energy Bioeffects Laboratory.

Belleau, E., Mariano, N., Novar, B. J., Chung, J., & Kuyk, T. K. (2015). *Joint helmet mounted cueing system visor laser eye protection ground and flight testing*. (AFRL-RH-FS-TR-2015-0051). Fort Sam Houston, TX 78234: Air Force Research Laboratory.

Bellenkes, A. H., Wickens, C. D., & Kramer, A. F. (1997). Visual scanning and pilot expertise: the role of attentional flexibility and mental model development. *Aviation, Space and Environmental Medicine, 68*(7), 569-579.

Biella, M. (2009). Pilot gaze performance in critical flight phases and during taxiing: Results from DLR-project MOSES. In K. KW, C. Bauer, M. Demendzic, S. Habershek, D. Jeloucan, B. Streit & C. Vogrincic (Eds.), *Aviation Psychology in Austria. Human Factors and Resources*.

Brockmeier, W. R., Kuyk, T. K., & Novar, B. J. (2014). *Optical quality testing of laser protective eyewear in RHDO: Equipment, methods and procedures*. (AFRL-RH-FS-TR-2014-0014). Fort Sam Houston, TX 78234: Air Force Research Laboratory.

Caldwell, J. A., Caldwell, J. L., Brown, D. L., & Smith, J. K. (2004). The effects of 37 hours of continuous wakefulness on the physiological arousal, cognitive performance, self-reported mood, and simulator flight performance of F-117A pilots. *Military Psychology, 16*(3), 163-181.

Cannavò, R., Conti, D., & Di Nuovo, A. (2016). Computer-aided assessment of aviation pilots attention: Design of an integrated test and its empirical validation. *Applied Computing and Informatics, 12*(1), 16-26.

Christ, R. E. (1975). Review and analysis of color coding research for visual displays. *Human Factors, 17*(6), 542-570.

Chuang, L. L., Nieuwenhuizen, F. M., & Bülthoff, H. H. (2013). A fixed-based flight simulator study: the interdependence of flight control performance and gaze efficiency. *Engineering Psychology and Cognitive Ergonomics, 8020*, 95-104.

Comstock, J. R., & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behavior research.* (NASA Technical Memorandum No. 104174). National Aeronautics and Space Administration, Hampton, NY: Langley Research Center.

Crognale, M. A., & Krebs, W. H. (2008). *Helicopter pilot performance: Inadvertent flight into instrument meteorological conditions*. Paper presented at the Human Factors and Ergonomics Society Annual Meeting, Los Angeles, CA.

Diez, M., Boehm-Davis, D. A., Holt, R. W., Pinney, M. E., Hansberger, J. T., & Schoppek, W. (2001). *Tracking pilot interactions with flight management systems through eye movements.* Paper presented at the 11th International Symposium on Aviation Psychology, Columbus, OH.

Doyon-Poulin, P., Ouellette, B., & Robert, J.-M. (2014). *Effects of visual clutter on pilot workload, flight performance and gaze pattern.* Paper presented at the International Conference on Human-Computer Interaction in Aerospace, Santa Clara, CA.

Dykes, J. R., Garcia, P. V., Maier, D. A., McLin, L. N., Ghani, N., Schmeisser, E. T., & Harrington, K. (2004). *Quantifying the effects of haze in laser eye protection on regan contrast letter acuity*. (AFRL-HE-BR-TR-2004-0055). Brooks Air Force Base, TX 78235: Air Force Research Laboratory.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.

Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998). *A comparative analysis of SAGAT and SART for evaluations of situation awareness.* Paper presented at the Human Factors and Ergonomics Society Annual Meeting, Los Angeles, CA.

Ford, M., & Munro, I. (2011). *Assessing pliot performance*. Paper presented at the PACDEFF Annual Conference, Queenstown, New Zealand.

Gaska, J. P., Wright, S. T., Winterbottom, M. D., & Hadley, S. C. (2016). Color vision and performance on color-coded cockpit displays. *Aerospace Medicine and Human Performance, 87*(11), 921-927.

Ginsburg, A. P., Easterly, J., & Evans, D. W. (1983). *Contrast sensitivity predicts target detection field performance of pilots.* Paper presented at the Human Factors Society Annual Meeting, Los Angeles, CA.

Ginsburg, A. P., Evans, D. W., Sekule, R., & Harp, S. A. (1982). Contrast sensitivity predicts pilots'performance in aircraft simulators. *Optometry and Vision Science, 59*(1), 105-108.

Handford, M. (1997a). *Where's Waldo now?* Somerville, MA: Candlewick Press.

Handford, M. (1997b). *Where's Waldo? The wonder book*. Somerville, MA: Candlewick Press.

Haslbeck, A., Kirchner, P., Schubert, E., & Bengler, K. (2014). *A flight simulator study to evaluate manual flying skills of airline pilots.* Paper presented at the Human Factors and Ergonomics Society Annual Meeting, Los Angeles, CA.

Haslbeck, A., Schubert, E., Gontar, P., & Bengler, K. (2012). The relationship between pilots' manual flying skills and their visual behavior: a flight simulator study using eye tracking. In W. Karwowski (Ed.), *Advances in Human Aspects of Aviation* (pp. 561-568). Boca Raton, FL: CRC Press.

Kasarskis, P., Stehwien, J., Hickox, J., Aretz, A., & Wickens, C. (2001). *Comparison of expert and novice scan behaviors during VFR flight.* Paper presented at the 11th International Symposium on Aviation Psychology, Columbus, OH.

Khan, M. J., Rossi, M., Heath, B., Ali, S. F., & Ward, M. (2006). *An experimental study of the effect of out-of-the-window cues on training novice pilots on a flight simulator*. Paper presented at the Human Factors and Ergonomics Society Annual Meeting, Los Angeles, CA.

Kirby, C. E., Kennedy, Q., & Yang, J. H. (2013). *An analysis of helicopter pilot scan techniques while flying at low altitudes and high speed.* Paper presented at the AIAA Atmospheric Flight Mechanics Conference, Boston, MA.

Kirby, C. E., Kennedy, Q., & Yang, J. H. (2014). Helicopter pilot scan techniques during low-altitude high-speed flight. *Aviation, Space, and Environmental Medicine, 85*(7), 740-744.

Kruk, R., Regan, D., Beverley, K., & Longridge, T. (1983). Flying performance on the advanced simulator for pilot training and laboratory tests of vision. *Human Factors, 25*(4), 457-466.

Kumagai, J. K., Williams, S., & Kline, D. (2005). *Vision standards for aircrew: Visual acuity for pilots*. (CR 2005-142). Toronto, Canada: DRDC Toronto.

Kuyk, T. K., Engler, S., Brockmeier, W. R., Kumru, S. S., & Mariano, N. (2013). *The effects of daytime laser eye protection for fixed and rotary wing aircraft on visual function*. (AFRL-RH-FS-TR-2013-0024). Fort Sam Houston, TX 78234: Air Force Research Laboratory.

Kuyk, T. K., Engler, S., Garcia, P. V., Smith, P. A., Novar, B. J., & Putnam, C. M. (2016). *Nighttime visor laser eye protection: visual function testing*. (AFRL-RH-FS-TR-2016-0013). Fort Sam Houston, TX 78234: Air Force Research Laboratory.

Kuyk, T. K., Smith, P. A., Engler, S., Garcia, P. V., Brockmeier, W. R., Novar, B. J., . . . McLin, L. N. (2013). *The effects of scattered light from optical components on visual function*. (AFRL-RH-FS-TR-2016-0018). Fort Sam Houston, TX 78234: Air Force Research Laboratory.

LaFrance, M., Kent, J. F., Foutch, B. K., Miller, M., & Kuyk, T. K. (2009). *Light transmission requirements for nighttime laser eye protection: preliminary findings*. (AFRL-HE-BR-TR-2009-0041). Brooks City Base, TX 78235: Air Force Research Laboratory.

Lee, K. (2010). *Effects of flight factors on pilot performance, workload, and stress at final approach to landing phase of flight.* (Doctor of Philosophy), University of Central Florida.

Lefrancois, O., Matton, N., Gourinat, Y., Peysakhovich, V., & Causse, M. (2016). *The role of pilots' monitoring strategies in flight performance.* Paper presented at the European Association for Aviation Psychology Conference EAAP32, Cascais, Portugal.

Leland, R., Rogers, R. O., Boquet, A., & Glaser, S. (2009). *An experiment to evaluate transfer of upset-recovery training conducted using two different flight simulation devices*. (DOT/FAA/AM-09/17). Washington, DC: Federal Aviation Adminstration.

Lindseth, P. D., Lindseth, G. N., Petros, T. V., Jensen, W. C., & Caspers, J. (2013). Effects of hydration on cognitive function of pilots. *Military Medicine, 178*(7), 792-798.

Lopez, N., Previc, F. H., Fischer, J., Heitz, R. P., & Engle, R. W. (2012). Effects of sleep deprivation on cognitive performance by United States Air Force pilots. *Journal of Applied Research in Memory and Cognition, 1*(1), 27-33.

Luder, C. B., & Barber, P. J. (1984). Redundant color coding on airborne CRT displays. *Human Factors, 26*(1), 19-32.

Maier, D. A., Brockmeier, W. R., Garcia, P. V., Salcedo, N., Kuyk, T. K., Dykes, J. R., & DeVilbiss, C. A. (2007). *Physical and Psychophysical Evaluations of ALEP ATD Laser Eye Protection Spectacles.* (AFRL-RH-BR-TR-2005-0136). Brooks City Base, TX 78235: Air Force Research Laboratory.

Martin, W. L., Murray, P. S., & Bates, P. R. (2012). *The effects of startle on pilots during critical events: A case study analysis.* Paper presented at the 30th EAAP Conference: Aviation Psychology & Applied Human Factors, Villasimius, Italy.

Martinsen, G., Havig, P., Dykes, J., Kuyk, T., & McLin, L. (2007). Night vision goggles, laser eye protection, and cockpit displays. *Proceedings of the SPIE, 6557*, 7.

Mavin, T., & Dall'Alba, G. (2010). *A model for integrating technical skills and NTS in assessing pilots' performance.* Paper presented at the 9th International Symposium of the Australian Aviation Psychology Association, Sydney, Australia.

Moacdieh, N. M., Prinet, J. C., & Sarter, N. B. (2013). *Effects of modern primary flight display clutter: Evidence from performance and eye tracking data.* Paper presented at the Human Factors and Ergonomics Society annual meeting, Los Angeles, CA.

Moacdieh, N. M., & Sarter, N. B. (2012). *Eye tracking metrics: A toolbox for assessing the effects of clutter on attention allocation.* Paper presented at the Human Factors and Ergonomics Society Annual Meeting, Los Angeles, CA.

Moroney, W. F., Biers, D. W., & Eggemeier, F. T. (1995). Some measurement and methodological considerations in the application of subjective workload measurement techniques. *The International Journal of Aviation Psychology, 5*(1), 87-106.

Mumenthaler, M. S., Yesavage, J. A., Taylor, J. L., O'Hara, R., Friedman, L., Lee, H., & Kraemer, H. C. (2003). Psychoactive drugs and pilot performance: a comparison of nicotine, donepezil, and alcohol effects. *Neuropsychopharmacology, 28*(7), 1366-1373.

Novar, B. J., Beilby, J., Zucker, A., Constable, W., Engler, S., Kuyk, T. K., & Smith, P. A. (2015). *Daytime visor laser eye protection: ground and flight testing*. (AFRL-RH-FS-TR-2015-0018). Fort Sam Houston, TX 78234: Air Force Research Laboratory.

Olmos, O., Wickens, C. D., & Chudy, A. (2000). Tactical displays for combat awareness: An examination of dimensionality and frame of reference concepts and the application of cognitive engineering. *The International Journal of Aviation Psychology, 10*(3), 247-271.

Post, D. L., Geiselman, E. E., & Goodyear, C. D. (1999). Benefits of color coding weapons symbology for an airborne helmet-mounted display. *Human Factors, 41*(4), 515-523.

Putnam, C. M., Goettl, B. P., Novar, B. J., Kuyk, T. K., & Smith, P. A. (2017). *Daytime visor laser eye protection: addendum - ground testing of modified prototypes*. (AFRL-RH-FS-TR-2017-0007). Fort Sam Houston, TX 78234: Air Force Research Laboratory.

Putnam, C. M., Novar, B. J., Goettl, B. P., Kuyk, T. K., Smith, P. A., & Engler, S. (2017). *Nighttime visor laser eye protection: ground and flight testing.* (AFRL-RH-FS-TR-2017-0007). Fort Sam Houston, TX 78234: Air Force Research Laboratory.

Reddix, M., Williams, H., Kirkendall, C., Eggan, S., Gao, H., Wells, W., & O'Donnell, K. (2014). *Assessment of color vision screening tests for U.S. Navy special duty occupations*. Paper presented at the Aerospace Medical Association 85th Annual Meeting, San Diego, CA.

Russo, M. B., Sing, H., Santiago, S., Kendall, A. P., Johnson, D., Thorne, D., . . . Redmond, D. (2004). Visual neglect: occurrence and patterns in pilots in a simulated overnight flight. *Aviation, Space, and Environmental Medicine, 75*(4), 323-332.

Salud, E. (2013). *Developing a library of display effects on pilot performance: Methods, meta-analyses, and performance estimates.* (Master of Science), San Jose State University.

Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data. *Human Factors, 49*(3), 347-357.

Schmeisser, E. T., Dykes, J. R., Ghani, N., Garcia, P. V., Maier, D. A., & Brockmeier, W. R. (1999). *Wardove laser eye protection (LEP) optical and psychophysical evaluations*. (AFRL-HE-BR-TR-1999-0223). Brooks Air Force Base, TX 78235: Air Force Research Laboratory.

Schmeisser, E. T., Maier, D. A., Freeman, D. A., & Brockmeier, W. R. (2005). *A psychophysical assesment of visibility in the cockpit: Laser eye protection, on-axis glare, and HUD legibility*. (AFRL-HE-BR-TR-2005-0061). Brooks Air Force Base, TX 78235: Air Force Research Laboratory.

Strater, L., & Endsley, M. (2000). *SAGAT: a situation awareness measurement tool for commercial airline pilots.* Paper presented at the First Human Performance, Situation Awareness, and Automation: User-Centered Design for the New Millennium Conference, Savannah, GA.

Sullivan, J., Yang, J. H., Day, M., & Kennedy, Q. (2011). Training simulation for helicopter navigation by characterizing visual scan patterns. *Aviation, Space, and Environmental Medicine, 82*(9), 871-878.

Taylor, J. L., Kennedy, Q., Noda, A., & Yesavage, J. A. (2007). Pilot age and expertise predict flight simulator performance A 3-year longitudinal study. *Neurology, 68*(9), 648-654.

Taylor, J. L., O'Hara, R., Mumenthaler, M. S., & Yesavage, J. A. (2000). Relationship of CogScreen-AE to flight simulator performance and pilot age. *Aviation, Space and Environmental Medicine, 7*, 373-380.

Tolton, R. G. (2014). *Relationship of individual pilot factors to simulated flight performance.* (Master of Health Sciences), University of Otago.

Wei, H., Zhuang, D., Wanyan, X., & Wang, Q. (2013). An experimental analysis of situation awareness for cockpit display interface evaluation based on flight simulation. *Chinese Journal of Aeronautics, 26*(4), 884-889.

Yang, J. H., Kennedy, Q., Sullivan, J., & Fricker, R. D. (2013). Pilot performance: assessing how scan patterns & navigational assessments vary by flight expertise. *Aviation, Space, and Environmental Medicine, 84*(2), 116-124.

Yesavage, J. A., Mumenthaler, M. S., Taylor, J. L., Friedman, L., O'Hara, R., Sheikh, J., . . . Whitehouse, P. J. (2002). Donepezil and flight simulator performance: effects on retention of complex skills. *Neurology, 59*(1), 123-125.

Yesavage, J. A., Taylor, J. L., Mumenthaler, M. S., Noda, A., & O'Hara, R. (1999). Relationship of age and simulated flight performance. *Journal of the American Geriatrics Society, 47*(7), 819-823.

Yu, C.-S., Wang, E. M.-Y., Li, W.-C., & Braithwaite, G. (2014). Pilots' visual scan patterns and situation awareness in flight operations. *Aviation, Space, and Environmental Medicine, 85*(7), 708-714.

Yu, C.-S., Wang, E. M.-Y., Li, W.-C., Braithwaite, G., & Greaves, M. (2016). Pilots' visual scan patterns and attention distribution during the pursuit of a dynamic target. *Aerospace Medicine and Human Performance, 87*(1), 40-47.

Zárate, D. (2012). *The effects of expertise and information location on change blindness detection within an aviation domain.* (Master of Science), Embry-Riddle Aeronautical University.